Cuce-H08945-11-1CP1116

THE JOURNAL OF
HILOSOPHY

05|9|2001
CALCUTTA UNIVERSITY LIBRARY
ALIPUR CAMPUS

VOLUME XCVIII, NUMBER 7
JULY 2001

11

page
331 *On Putnam and His Models* Timothy Bays
351 *On Sense and Reflexivity* John Justice

COMMENTS AND CRITICISM
365 *On Multiple Realization and the Special Sciences* Alex Rosenberg

BOOK REVIEWS
374 Russell Hardin *Liberalism, Constitutionalism, and Democracy* Robert E Goodin

379 NEW BOOKS

ublished by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J E WOODBRIDGE AND WENDELL T BUSH

Purpose To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines

Editors Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane *Editor Emeritus* Sidney Morgenbesser *Consulting Editors* David Albert, John Collins, James T Higginbotham, Charles D Parsons, Achille C Varzi *Managing Editor* Michael Kelly

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc *President,* Arthur C Danto, *Vice President,* Akeel Bilgrami, *Secretary,* Daniel Shapiro, *Treasurer,* Barbara Gimbel, *Other Trustees* Leigh S Cauman, Kent Greenawalt, Michael J Mooney, Lynn Nesbit

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027 FAX (212) 932-3721

THE JOURNAL OF PHILOSOPHY

2001

SUBSCRIPTIONS (12 issues)

Individuals	$35 00
Libraries and Institutions	$75 00
Students, retired/unemployed philosophers	$20 00
Postage outside the U S	$15 00

Payments only in U S currency on a U S bank All back volumes and separate issues available back to 1904 Please inquire for price lists, shipping charges, and discounts on back orders Please inquire for advertising rates, ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977, typeset and printed by Cadmus Journal Services, Lancaster and Akron, PA

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027 (212) 866-1742

CALCUTTA UNIVERSITY LIBRARY
ALIPORE CAMPUS
Acc. K-P-1116
Date 13.09.02

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U S Copyright Law This consent does not extend to any other kinds of copying More information on our terms may be obtained by consulting our January issue or by writing to us

© Copyright 2001 by the Journal of Philosophy, Inc

ISSN 0022-362X
POSTMASTER Periodical postage paid at New York, NY, and other mailing offices

THE JOURNAL OF
HILOSOPHY

UNIVERSITY

VOLUME XCVIII, NUMBER 8
AUGUST 2001

page
389 *Neutrality and Responsibility* Steven Wall
411 *Knowing How* Jason Stanley and Timothy Williamson

ublished by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J E WOODBRIDGE AND WENDELL T BUSH

Purpose To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines

Editors Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane, Achille C Varzi *Editor Emeritus* Sidney Morgenbesser *Consulting Editors* David Albert, John Collins, James T Higginbotham, Charles D Parsons *Managing Editor* Michael Kelly

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc *President,* Arthur C Danto, *Vice President,* Akeel Bilgrami; *Secretary,* Daniel Shapiro, *Treasurer,* Barbara Gimbel, *Other Trustees* Leigh S Cauman, Kent Greenawalt, Michael J Mooney, Lynn Nesbit

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027 FAX (212) 932-3721

You may also visit our website at wwwjournalofphilosophy org

THE JOURNAL OF PHILOSOPHY

2001

SUBSCRIPTIONS (12 issues)

Individuals	$35 00
Libraries and Institutions	$75 00
Students, retired/unemployed philosophers	$20 00
Postage outside the U S	$15 00

Payments only in U S currency on a U S bank All back volumes and separate issues available back to 1904 Please inquire for price lists, shipping charges, and discounts on back orders Please inquire for advertising rates, ad space is limited, so ad reservations are taken on a first come, first served basis

Published monthly as of January 1977, typeset and printed by Cadmus Journal Services, Lancaster and Akron, PA

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027 (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U S Copyright Law This consent does not extend to any other kinds of copying More information on our terms may be obtained by consulting our January issue or by writing to us

© Copyright 2001 by the Journal of Philosophy, Inc ISSN 0022-362X
POSTMASTER Periodical postage paid at New York, NY, and other mailing offices

THE JOURNAL OF
HILOSOPHY

VOLUME XCVIII, NUMBER 9
SEPTEMBER 2001

page
445 *The Groundless Normativity of Instrumental Rationality* Donald C Hubin
469 *Of Colors, Kestrels, Caterpillars, and Leaves* Peter Bradley and Michael Tye

REVIEW ESSAYS
488 *Recent Work in Philosophy of Mathematics* Jamie Tappenden

498 NOTES AND NEWS

blished by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J E WOODBRIDGE AND WENDELL T BUSH

Purpose To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines

Editors Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane, Achille C Varzi *Editor Emeritus* Sidney Morgenbesser *Consulting Editors* David Albert, John Collins, James T Higginbotham, Charles D Parsons *Managing Editor* Michael Kelly

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc *President,* Arthur C Danto, *Vice President,* Akeel Bilgrami, *Secretary,* Daniel Shapiro, *Treasurer,* Barbara Gimbel, *Other Trustees* Leigh S Cauman, Kent Greenawalt, Michael J Mooney, Lynn Nesbit

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027 FAX (212) 932-3721
You may also visit our website at www journalofphilosophy org

THE JOURNAL OF PHILOSOPHY

2001

SUBSCRIPTIONS (12 issues)

Individuals	$35 00
Libraries and Institutions	$75 00
Students, retired/unemployed philosophers	$20 00
Postage outside the U S	$15 00

Payments only in U S currency on a U S bank All back volumes and separate issues available back to 1904 Please inquire for price lists, shipping charges, and discounts on back orders Please inquire for advertising rates, ad space is limited, so ad reservations are taken on a first come, first served basis

Published monthly as of January 1977, typeset and printed by Cadmus Journal Services, Lancaster and Akron, PA

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027 (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U S Copyright Law This consent does not extend to any other kinds of copying More information on our terms may be obtained by consulting our January issue or by writing to us

© Copyright 2001 by the Journal of Philosophy, Inc ISSN 0022-362X
POSTMASTER Periodical postage paid at New York, NY, and other mailing offices

THE JOURNAL OF HILOSOPHY

UNIVERSITY LIBRARY
ALIPUR
CAMPUS

VOLUME XCVIII, NUMBER 10
OCTOBER 2001

page
499 *Coining Terms in the Language of Thought Innateness, Emergence, and the Lot of Cummins's Argument Against the Causal Theory of Mental Content* Robert D Rupert
531 *Alter Egos and Their Names* David Pitt

553 NEW BOOKS ANTHOLOGIES

ublished by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J E WOODBRIDGE AND WENDELL T BUSH

Purpose To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines

Editors Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane, Achille C Varzi *Editor Emeritus* Sidney Morgenbesser *Consulting Editors* David Albert, John Collins, James T Higginbotham, Charles D Parsons *Managing Editor* Michael Kelly

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc *President,* Arthur C Danto, *Vice President,* Akeel Bilgrami, *Secretary,* Daniel Shapiro, *Treasurer,* Barbara Gimbel, *Other Trustees* Leigh S Cauman, Kent Greenawalt, Michael J Mooney, Lynn Nesbit

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027 FAX (212) 932-3721

You may also visit our website at www journalofphilosophy org

THE JOURNAL OF PHILOSOPHY

2001

SUBSCRIPTIONS (12 issues)

Individuals	$35 00
Libraries and Institutions	$75 00
Students, retired/unemployed philosophers	$20 00
Postage outside the U S	$15 00

Payments only in U S currency on a U S bank All back volumes and separate issues available back to 1904 Please inquire for price lists, shipping charges, and discounts on back orders Please inquire for advertising rates, ad space is limited, so ad reservations are taken on a first come, first served basis

Published monthly as of January 1977, typeset and printed by Cadmus Journal Services, Lancaster and Akron, PA

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027 (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U S Copyright Law This consent does not extend to any other kinds of copying More information on our terms may be obtained by consulting our January issue or by writing to us

© Copyright 2001 by the Journal of Philosophy, Inc ISSN 0022-362X

POSTMASTER Periodical postage paid at New York, NY, and other mailing offices

7.1.02

THE JOURNAL OF PHILOSOPHY

VOLUME XCVIII, NUMBER 11
NOVEMBER 2001

page
555 *Bayesian Epistemology and Epistemic Conditionals: On the Status of the Export-Import Laws* Horacio Arló-Costa

COMMENTS AND CRITICISM
594 *When "Evidence Transcendence" Is Not Malign: A Reply to Crispin Wright* Hilary Putnam

BOOK REVIEWS
601 Michael Friedman: *A Parting of the Ways: Carnap, Cassirer, and Heidegger* Hans Sluga

612 NEW BOOKS: TRANSLATIONS

Published by The Journal of Philosophy, Inc.

THE JOURNAL OF PHILOSOPHY
FOUNDED BY FREDERICK J E WOODBRIDGE AND WENDELL T BUSH

Purpose To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines

Editors Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane, Achille C Varzi *Editor Emeritus* Sidney Morgenbesser *Consulting Editors* David Albert, John Collins, James T Higginbotham, Charles D Parsons *Managing Editor* Michael Kelly

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc *President,* Arthur C Danto, *Vice President,* Akeel Bilgrami, *Secretary,* Daniel Shapiro, *Treasurer,* Barbara Gimbel, *Other Trustees* Leigh S Cauman, Kent Greenawalt, Michael J Mooney, Lynn Nesbit

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027 FAX (212) 932-3721
You may also visit our website at www journalofphilosophy org

THE JOURNAL OF PHILOSOPHY
2001
SUBSCRIPTIONS (12 issues)

Individuals	$35 00
Libraries and Institutions	$75 00
Students, retired/unemployed philosophers	$20 00
Postage outside the U S	$15 00

Payments only in U S currency on a U S bank All back volumes and separate issues available back to 1904 Please inquire for price lists, shipping charges, and discounts on back orders Please inquire for advertising rates, ad space is limited, so ad reservations are taken on a first come, first served basis

Published monthly as of January 1977, typeset and printed by Cadmus Journal Services, Lancaster and Akron, PA

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027 (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U S Copyright Law This consent does not extend to any other kinds of copying More information on our terms may be obtained by consulting our January issue or by writing to us

© Copyright 2001 by the Journal of Philosophy, Inc
ISSN 0022-362X
POSTMASTER Periodical postage paid at New York, NY, and other mailing offices

THE JOURNAL OF PHILOSOPHY

25|11|02

VOLUME XCVIII, NUMBER 12
DECEMBER 2001

page
613 *Two Dogmas of Davidsonian Semantics* Max Kolbel
636 *Self-defeating Symbolism in Politics*
Guido Pincione and Fernando R Tesón

BOOK REVIEWS
653 John Foster *The Nature of Perception* Frank Jackson

658 INDEX TO VOLUME XCVIII

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J E WOODBRIDGE AND WENDELL T BUSH

Purpose To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines

Editors Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane, Achille C Varzi *Editor Emeritus* Sidney Morgenbesser *Consulting Editors* David Albert, John Collins, James T Higginbotham, Charles D Parsons *Managing Editor* Michael Kelly

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc *President,* Arthur C Danto, *Vice President,* Akeel Bilgrami, *Secretary,* Daniel Shapiro, *Treasurer,* Barbara Gimbel, *Other Trustees* Leigh S Cauman, Kent Greenawalt, Michael J Mooney, Lynn Nesbit

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027 FAX (212) 932-3721

You may also visit our website at www journalofphilosophy org

THE JOURNAL OF PHILOSOPHY

2001

SUBSCRIPTIONS (12 issues)

Individuals	$35 00
Libraries and Institutions	$75 00
Students, retired/unemployed philosophers	$20 00
Postage outside the U S	$15 00

Payments only in U S currency on a U S bank All back volumes and separate issues available back to 1904 Please inquire for price lists, shipping charges, and discounts on back orders Please inquire for advertising rates, ad space is limited, so ad reservations are taken on a first come, first served basis

Published monthly as of January 1977, typeset and printed by Cadmus Journal Services, Lancaster and Akron, PA

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027 (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U S Copyright Law This consent does not extend to any other kinds of copying More information on our terms may be obtained by consulting our January issue or by writing to us

© Copyright 2001 by the Journal of Philosophy, Inc

ISSN 0022–362X
POSTMASTER Periodical postage paid at New York, NY, and other mailing offices

THE JOURNAL OF PHILOSOPHY
VOLUME XCVIII, NO 7, JULY 2001

5191

Philosof

ON PUTNAM [illegible] S MODELS*

> It is not my claim that the 'Lowenheim-Skolem paradox' is an antinomy *in formal logic*, but I shall argue that it *is* an antinomy, or something close to it, in *philosophy of language* Moreover, I shall argue that the resolution of the antinomy—the only resolution that I myself can see as making sense—has profound implications for the great metaphysical dispute about realism which has always been the central dispute in the philosophy of language [1]

For the past twenty years or so, Hilary Putnam has advanced a family of arguments which go under the patronymic "the model-theoretic argument against realism " These arguments purport to show that basic theorems of model theory entail that realistic accounts of truth and reference are untenable and that, as a result, realistic metaphysics is incoherent On the basis of these arguments, Putnam urges philosophers to abandon traditional realism—or *metaphysical realism*, as he likes to call it—and to adopt his own, new-and-improved, *internal realism* instead

Here, I examine one version of Putnam's argument—a version closely related to the traditional Löwenheim-Skolem paradox I defend three claims about this argument First, I argue that a key step in Putnam's argument rests on a mathematical mistake, and I discuss some of the philosophical ramifications of this mistake Second, I argue that, even if Putnam could get his mathematics to work, his argument would still fail on purely philosophical grounds Third, I argue that Putnam's mathematical mistakes and his philosophical

---

* This paper has benefited from discussions with members of the philosophy departments at the University of California/Los Angeles, Yale University, the University of Notre Dame, and Northwestern University I am particularly grateful to Tony Martin and Yiannis Moschovakis for helpful comments on earlier drafts

[1] Hilary Putnam, "Models and Reality," in *Realism and Reason* (New York Cambridge, 1983), pp 1-25 Hereafter MR

0022-362X/01/9807/331-50

mistakes are surprisingly closely related. At the end of the day, I conclude that realists have little to fear from Putnam and his models.

I. PUTNAM'S ARGUMENT

The version of Putnam's argument that I want to consider occurs in the first fifteen pages of a paper entitled "Models and Reality" (MR). The paper begins with a short discussion of the Lowenheim-Skolem paradox and then transforms this paradox in two ways. For expository convenience, I shall follow this same progression in laying out Putnam's argument.

Begin with a standard, first-order axiomatization of set theory—say, *Zermelo-Fraenkel set theory with the axiom of choice* (ZFC). On the assumption that this axiomatization has a model, the Lowenheim-Skolem theorems ensure that it has a countable model. Call this model M. Now, because ZFC $\vdash \exists x$ '*x is uncountable*', there must be some $m \in M$ such that

$$M \vDash \text{'}m \textit{ is uncountable}\text{'}$$

Since M itself is only countable, however, there are only countably many $n \in M$ such that $M \vDash n \in m$. Thus, cardinality seems to be relative: from one perspective, m seems to be uncountable, while from another perspective, m is clearly countable.

Putnam begins his analysis of this paradox by noting that, whatever it might show about the countability or uncountability of m, the paradox highlights the fact that ZFC has *different* models. Further, these models interpret many of the central definitions of classical set theory in structurally different ways. Some models satisfy 'm is uncountable' only if m really is uncountable, while other models (like the M above) satisfy 'm is uncountable' even when m is really countable. Similarly, some models satisfy sentences like 'm is finite' or 'm is the power set of n' if and only if m really is finite or really is the power set of n, while other models satisfy these sentences under (quite) different circumstances.

Skolem's paradox shows, therefore, that, if the meanings of phrases like 'is countable' or 'is finite' are fixed only by the model theory of first-order set theory, then these phrases are semantically indeterminate. To use Putnam's terminology, the paradox shows that it is impossible to pin down the "intended interpretation" of set-theoretic language using only (first-order) axioms. Further, Putnam clearly thinks that axioms are all we have to go on here: "but if *axioms* cannot capture the 'intuitive notion of set' what possibly could" (MR, 3)? As a result, Putnam concludes that even traditional formulations of Skolem's paradox uncover a genuine, and a fairly general, indeterminacy in our use of set-theoretic language.

So much, then, for the traditional Lowenheim-Skolem paradox. Putnam's first transformation of this paradox involves parlaying the general indeterminacy of set-theoretic language into more specific indeterminacies in the truth values of set-theoretic sentences. Putnam writes: "If I am right, then the 'relativity of set-theoretic notions' extends to a *relativity of the truth value of '$V = L$'* (and, by similar arguments, of the axiom of choice and the continuum hypothesis as well)" (MR, 7-8). His idea here is simple. It is a result of modern set theory that $V = L$ is "independent" of ZFC—that, on the assumption that ZFC itself is consistent, so are ZFC + $V = L$ and ZFC + $V \neq L$. Hence, if the intended model of set theory is fixed *only* by the axioms of ZFC—and if there is, in fact, some such intended model—then there is an intended model in which $V = L$ is true and another in which $V = L$ is false.

Putnam's second transformation involves arguing that, to the extent that axioms fail to ensure the semantic determinacy of set-theoretic language, physical science will not pick up the slack. In particular, physical science will not fix a unique "intended interpretation" for the vocabulary of set theory, nor will it restrict such interpretations so as to eliminate the indeterminacy in truth values of set-theoretic sentences. Now, at first glance, this second transformation may seem somewhat unmotivated. How, after all, *could* physical science affect the interpretation of basic set theory? Later glances reveal, however, that Putnam has something to worry about here, and that his worries are closely related to the project of showing that $V = L$ has an indeterminate truth value.

Suppose we have a machine that takes a measurement—of something, it does not matter what—every three or four seconds. Suppose also that this machine gives a reading of 1 or 0 depending on the results of its measurements, and suppose finally that this machine manages to run for an *infinite* period of time and (so) produces an infinite sequence of measurements. In theory, the sequence of ones and zeros which result from these measurements could "code up" a nonconstructible set—that is, a set that lives in V but not in L. In this case, it might seem like nature itself manages to falsify the hypothesis that $V = L$. This possibility, together with Putnam's desire to make '$V = L$' come out indeterminate, explains why he thinks he needs an explicit argument concerning the impact of physical science on the interpretation of set theory.

Putnam's explicit argument comes in two parts. The first centers on the following theorem which he states and proves (MR, 6-7).

Theorem 1. *ZF plus $V = L$ has an ω-model which contains any given countable set of real numbers.*

Given this theorem, Putnam argues as follows. Let OP be a countable collection of real numbers which codes up all of the measurements human beings will ever make (including measurements made by machines like those discussed above). By theorem 1, there is a model of ZF + $V = L$ which contains OP (or, at least, a formal analog of OP). Since this model satisfies ZFC it must be an "intended model," and since it both satisfies $V = L$ and contains OP, it takes care of the measurement problem discussed in the last paragraph.

Next, Putnam argues that this measurement problem is the *only* problem with physical science that *needs* to be taken care of. He writes:

> Now, suppose we formalize *the entire language of science* within the set theory ZF *plus* $V = L$. Any model for ZF which contains an abstract set isomorphic to OP can be extended to a model for this formalized language of science which is *standard with respect to OP*, hence we can find a model *for the entire language of science* which satisfies '*everything is constructible*' and which assigns the correct value to all physical magnitudes (MR, 7).

Thus, as long as the only constraints on the interpretation of set-theoretic vocabulary come from the formal structure of our scientific theories (including the explicit axioms of our set theory) and from the physical measurements we happen to make, there will be *an* interpretation of set theory on which $V = L$ comes out true. On the flip side, Putnam assumes that there is *some* interpretation of set theory—again, an interpretation compatible with the rest of our scientific theories and with all the physical measurements we might ever happen to make—on which $V = L$ comes out false.[2]

This, then, gives us an initial premise in Putnam's model-theoretic argument. Let the phrase 'theoretical constraints' refer to the set of sentences that constitutes our best theory of the world—that is, our best physical theory together with our best axioms for set theory—and let 'operational constraints' refer to all the measurements we might ever happen to make. Then we have

---

[2] Putnam makes this assumption for two reasons. From a dialectical perspective, he takes himself to be arguing against Kurt Godel. Godel thought there *was* a unique "intended interpretation" of set theory and that $V = L$ comes out false on this interpretation. Hence, the existence of intended models satisfying $V \neq L$ is not really at issue in this context.

From a somewhat deeper perspective, it is fairly easy to start with a model that does satisfy $V = L$ and expand it to one that does *not* (and to preserve nice properties like "being an ω-model" in the process). It is substantially more difficult to start with a model that does *not* satisfy $V = L$ and expand it to one that does. This is undoubtedly the main reason Putnam puts so much effort into obtaining a model that satisfies $V = L$ and so little into obtaining one that satisfies $V \neq L$.

(1) Theoretical and operational constraints do not fix a unique "intended interpretation" for the language of set theory.

Further, Putnam clearly thinks that theoretical and operational constraints are the only factors that *could* fix an "intended interpretation" for set theory. His reasons for thinking this will be discussed in section iii. For now, let me simply lay out the remainder of his argument.

(2) Nothing other than theoretical and operational constraints could fix a unique "intended interpretation" for the language of set theory. So,

(3) There is no unique "intended interpretation" for the language of set theory.

Finally, because different, equally "intended," interpretations of set theory disagree on the truth value of sentences like $V = L$, there simply are no determinate truth values for such sentence. They are, in Putnam's words, "just true in some intended models and false in others" (MR, 5).

This, therefore, gives us the overall structure of Putnam's argument. The goal is to show that set-theoretic language is semantically indeterminate. To reach this goal, we first note that the axioms of set theory do not determine a unique interpretation for set-theoretic language. Next, we observe that throwing in scientific information—that is, the physical theories and measurements that fill out Putnam's "theoretical and operational constraints"—will not improve the situation. Finally, we notice that the different "intended interpretations" that are compatible with our "theoretical and operational constraints" disagree on the truth value of sentences like $V = L$, and we conclude that these truth values are, themselves, indeterminate.

II. THE MATHEMATICS OF PREMISE (1)

In this section, I slow down and examine more carefully the details of Putnam's mathematics. My main theses are straightforward. I argue (1) that Putnam's proof of theorem 1 is mistaken, (2) that this mistake cannot be "patched up" without weakening the overall force of Putnam's argument, and (3) that even the weakened version of Putnam's argument leaves the realist with significant problems.

To begin, consider Putnam's proof of theorem 1. The theorem says that, if X is a countable collection of real numbers, then there exists an ω-model, M, such that $M \models \mathrm{ZF} + V = L$, and M contains an "abstract copy" of the set X. Putnam's proof begins by noting that, in the special case in which we allow M to be countable, we can code both M and X by single reals In this case, the theorem can be formulated as a Π_2 sentence of the form (For every real s) (There is a real M) such that ($M, s,$) From here, Putnam argues as follows

> Consider this sentence *in the inner model* $V = L$ For every s *in the inner model*—i e , for every s in L—there is a model—namely L itself—which satisfies '$V = L$' and contains s By the downward Lowenheim-Skolem theorem, there is a countable submodel which is elementarily equivalent to L and contains s (Strictly speaking, we need here not just the downward Lowenheim-Skolem theorem, but the 'Skolem Hull' construction which is used to prove that theorem) By Godel's work, this countable submodel itself lies in L, and, as is easily verified, so does the real that codes it So, the above Π_2-sentence is true in the inner model $V = L$
>
> But Shoenfield has proved that Π_2-sentences are *absolute* if a Π_2-sentence is true in L, then it must be true in V So the above sentence is true in V (MR, 6)

Ironically, the problem with this "proof" involves Putnam's application of the Lowenheim-Skolem theorem The "short version" of the problem is simple the downward Lowenheim-Skolem theorem applies only to structures that have sets for their domains, and L—the structure to which Putnam applies the downward Lowenheim-Skolem theorem—does not have a set for its domain Hence, Putnam cannot (legitimately) use the downward Lowenheim-Skolem theorem to obtain "a countable submodel which is elementarily equivalent to L and contains [the real] s " Without this submodel, Putnam cannot ensure that L satisfies the Π_2 sentence that he needs it to satisfy hence, he cannot apply Shoenfield absoluteness to "reflect" this sentence up to V In the absence of Putnam's countable submodel, therefore, his overall proof simply collapses[3]

---

[3] The "long version" of this problem simply expands the "short version" with a fair bit of terminological clarification To begin, set theorists typically distinguish between *sets* and *proper classes* Roughly, sets are classes which are "small enough" to be members of other classes, while proper classes are classes that are "too big" to count as sets Examples of sets would include $\varnothing$, $\aleph_\omega$, and the power set of $\mathbb{N}$ Examples of proper classes would include the class of all sets, the class of all ordinals, and the class of all countable sets Most pertinently, the class of all constructible sets—that is, L—is a proper class

It is important here, for reasons which will become clear shortly, that proper classes are usually required to be *definable* That is, for some formula in the language of set theory, $\phi(x, y_1, \; , y_n)$, and some sequence, $a_1, \; , a_n$, of sets, we can consider the class of all sets b such that $\phi(b, a_1, \; , a_n)$ is true To insist that proper classes be definable is to insist that *all* classes be picked out by a formula of this kind

II 1 Five technical comments Before I discuss the philosophical ramifications of the failure of Putnam's proof, I want to make five technical comments about the failure itself First, Putnam's proof is not saved by his qualification "Strictly speaking, we need here not just the downward Lowenheim-Skolem theorem, but the 'Skolem Hull' construction which is used to prove that theorem " To be sure, the Skolem Hull construction allows us to prove so-called "reflection theorems" in which some *finite* collection of sentences is "reflected" from a proper class to a set That is, if we start with a proper class that "satisfies" some finite collection sentences, then the Skolem Hull construction lets us find a countable set that satisfies the same collection of sentences [4] But this construction works *only* when we try to reflect *finite* collections of sentences In particular, then, it does not allow Putnam to reflect the *infinite* collection, ZF + $V = L$

Second, there is nothing particularly surprising about the fact that Putnam's proof fails Leaving aside the details of this proof, consider just the *form* of theorem 1 for any countable set of reals X, there is an ω-model M, such that $M \models$ ZF + $V = L$ and $X \in M$ Now, since any

---

Second, we need to distinguish between two interpretations of the terms 'model' and 'satisfaction' In ordinary model theory, the term 'model' refers exclusively to structures which have *sets* for their domains In turn, the 'satisfaction' relation is defined as a relation *between sets*—that is, between the sets which constitute the domain (and relations) of a model and those which code up the formulas of our language Set theorists often use the terms 'model' and 'satisfaction' differently They speak of 'class models' when they want to refer to proper classes in which certain collections of sentences hold, and they use 'satisfaction' to refer to the fact that certain sentences become true when their quantifiers are relativized to a proper class—that is, explicitly relativized using the formula that defines the class in question (see above) It is in this latter sense, for instance, that Putnam refers to 'inner models'

For our purpose, the key fact concerning all this terminology is the following the Löwenheim-Skolem theorems apply to set models, but they do not apply to class models Hence, we cannot use the downward Löwenheim-Skolem theorem to find an elementary submodel of some class model of ZF In particular, Putnam cannot use the downward Lowenheim-Skolem theorem to find a (countable) model which is elementarily equivalent to L

Given this, Putnam faces a dilemma If he intends the term 'model' to refer to set models, then his proof goes wrong when he says "for every s in L—there is a model—namely L itself—which satisfies '$V = L$' " For, since L is not a model at all, it is not a model satisfying ZF + $V = L$ If, on the other hand, Putnam intends the term 'model' to refer both to set models and to class models, then he can legitimately call L a 'model' But even then, he cannot apply the downward Löwenheim-Skolem theorem to this model On any interpretation, then, Putnam's proof is mistaken However 'model' gets defined, his proof requires that we apply the downward Löwenheim-Skolem theorem to L, and, as we have seen, this is just not possible

[4] Here, we have an example of a *legitimate* shift between the two uses of 'satisfies' that I mentioned in the last footnote

model of ZF + $V = L$ is also a model of ZFC, theorem 1 entails that there is a model of ZFC And since this, in turn, entails that ZFC is consistent, theorem 1 also entails that ZFC is consistent

By Gödel's second incompleteness theorem, however, the consistency of ZFC *cannot* be proved within ZFC itself As a corollary, then, theorem 1 cannot be proved in ZFC Therefore, we should not be surprised to find that there is a mistake in Putnam's proof his proof must be mistaken, because theorem 1 *cannot* be proved using the set theory with which he is working Hence, unless Putnam is willing to adopt some stronger set theory, his overall argument is bound to fail[5]

Third, Putnam's proof is relatively easy to "patch up" If he is willing to adopt a stronger set theory, then he can salvage his theorem (and, for the most part, his proof) So, for instance, if Putnam were willing to accept the existence of inaccessible cardinals, then his proof could be reconstituted with only minor modifications[6] Similarly, if he were willing to extend ZFC with a collection of axioms governing the behavior of proper classes, then he could probably prove "full reflection theorems" that would allow him to obtain (elementary) submodels of proper classes like V and L, this would, once again, allow him to reconstitute the essentials of his original proof In either case, therefore, a slight strengthening of Putnam's background mathematics allows him to save his proof from the problem discussed above

Fourth, although these "patching strategies" allow Putnam to salvage his proof, they do very little toward salvaging his overall philosophical argument Very roughly, patchings of the type just mentioned leave his critics with two fairly obvious lines of response On the one hand, some philosophers might reject Putnam's argument simply because they reject his new mathematics Whereas ZFC is a relatively widely-accepted axiomatization of set theory, the extensions discussed above are less widely accepted Hence, a philosopher (or mathematician) who has reservations about inaccessible cardinals and/or strong class axioms might well reject Putnam's new argument just because it employs this extra mathematics Indeed, such a philosopher might even think that she has been given new *reasons* for rejecting Putnam's extra mathematics "if Putnam's new math generates arguments for semantic indeterminacy, then so much the worse for Putnam's new math "

---

[5] In conversation, several people have suggested that Putnam might not be working in ZFC in the first place—that is, that he might *already* be working in some stronger form of set theory But, while this suggestion helps to save his mathematics, it faces two major difficulties The first is textual Putnam's entire paper is *about* ZFC, and throughout the paper he writes as though ZFC is the *obvious* set theory in which to work Hence, it would be strange to find that his entire argument rests on an (utterly unmentioned) version of set theory which he keeps hidden in the background The second difficulty is philosophical for reasons which will be discussed shortly, working with a stronger set theory would not, in the long run, save Putnam's argument (although it *might*, in the short run, allow theorem 1 to be proved properly)

[6] To see this, let κ be the inaccessible cardinal in question At the point in Putnam's proof where he claims that L is a model for ZF + $V = L$ and that L contains the real s, he can simply argue that L_κ is a model for ZF + $V = L$ and that L_κ contains the real s Since L_κ *really is* a set model for ZF + $V = L$, Putnam can proceed to apply the downward Löwenheim-Skolem theorem to L_κ to obtain his desired model M From here, his proof goes as before

On the other hand, even philosophers who do accept Putnam's strenghtened mathematics—say, those who accept the axiom of inaccessible cardinals—have ample grounds for rejecting his overall argument In particular, they should reject Putnam's claim that his model *M*—the model guaranteed by theorem 1—satisfies "all theoretical constraints " Since theorem 1 does not guarantee that *M* satisfies the sentence 'There exists an inaccessible cardinal', *M* may not even satisfy the "theoretical constraints" imposed by set theory Contra Putnam, then, there is no reason to think that *M* provides an "intended interpretation" for set-theoretic language

Recall, here, the philosophical point of Putnam's theorem He wants a model of ZF + $V = L$ which "satisfies all theoretical constraints [and] all operational constraints as well" (MR, 7) His theorem is supposed to provide such a model the theory of the model takes care of "theoretical constraints" and the fact that s is a member of the model takes care of "operational constraints " With this model in hand, Putnam tries to argue that $V = L$ is true in the (or at least in *an*) intended model of set theory.

My point is simply this. if our working set theory is *stronger* than ZFC—because we accept inaccessible cardinals, or class axioms, or whatever else is needed to patch up Putnam's proof—then it is hard to see how his theorem accomplishes its goal Given that we accept more mathematics than ZFC, this new mathematics should count as part of our "theoretical constraints." Thus, it is not enough for Putnam to build a model that satisfies ZF + $V = L$, he needs a model that satisfies ZF + $V = L$ *plus whatever else we happen to have added to ZFC* Since Putnam's theorem does not, so far as we know, provide a model satisfying this *extended* theory, it does not do what he wants it to do

Fifth, this problem is intrinsic to the *kind* of argument Putnam wants to make. Returning again to the incompleteness theorem, I note that there is no way for him both (1) to use a particlar axiomatization of set theory (say, ZFC + XYZ) as his background set theory and (2) to prove the existence of a model satisfying ZFC + XYZ + $V = L$. Hence, *whatever* version of set theory Putnam winds up working in, he will be unable to tailor his overall argument so as to take care of the "theoretical constraints" this version imposes.

As a result, Putnam faces an *inescapable* dilemma. If he pitches his argument toward philosophers who accept less set theory than he himself does, then these philosophers will reject his argument simply because they reject the set theory used in proving Putnam's key theorem. If he pitches his argument toward philosophers who accept the same set theory that he does, then his argument cannot take care of *these philosophers'* "theoretical constraints." At the end of the day, therefore, *no one* has adequate grounds for accepting Putnam's model-theoretic argument.[7]

This, therefore, gives us a first, and essentially technical, response to Putnam's argument. His argument depends on a key theorem which he is not in a position to prove. Nor, for reasons relating to Gödel's second incompleteness theorem, can he place himself in such a position without jeopardizing the very philosophical point that his theorem is supposed to support. In short, the mathematical mistake which we discussed at the beginning of this section is one which cannot be fixed without undercutting the model-theoretic argument as a whole.

II.2. Two philosophical comments. How strong is this argument which I have just sketched? Unfortunately, much as I like the argument, I am afraid the answer is "not very." On the positive side, the argument shows that Putnam cannot *conclusively prove* that set theory is semantically indeterminate. That is, if we want Putnam to stand toe-to-toe with the realist and *prove* semantic antirealism, then my argument shows that he cannot do it.

On the negative side, Putnam has clearly given an argument that raises the *possibility* of semantic indeterminacy. Suppose that premise (2) in his argument is correct. Then the only way for set-theoretic language to be semantically determinate is for there to be a *unique* model that satisfies all our "theoretical and operational constraints." To the extent that alternative models *happen to exist*, set theory winds up being semantically indeterminate.

---

[7] Note that this argument is not specific to the search for a model satisfying $V = L$; it works equally well against attempts at finding models for ZFC ± CH or ZF ± C. Indeed, when framed as a simple incompleteness problem, it applies even to the "plain vanilla" version of Skolem's paradox which simply tries to find structurally dissimilar models for ZF.

In this context, the mere fact that Putnam cannot *prove* his central theorem should provide very little comfort to the realist If the technical response to his argument is all we have to go on—if, that is, we are willing to accept premise (2) and to rest our challenge to premise (1) solely on the considerations discussed above—then realism depends on the *mere hope* that Putnam's nonstandard models do not exist This is not much to stake a metaphysics on![8]

All that being said, this argument clearly depends on the assumption that premise (2) is true If it is not, then the existence of nonstandard models for our "theoretical and operational constraints" becomes considerably less troubling It is high time, therefore, that we turn and examine premise (2)

III JUST MORE THEORY I

Although premise (1) of Putnam's model-theoretic argument has attracted a great deal of attention in the literature—and although it is clearly the premise that gives Putnam's argument most of its philosophical "sex appeal"—it should be fairly clear that premise (2) is where the real philosophical action takes place For one thing, premise (1) only eliminates one method for fixing the intended interpretation of set theory, while premise (2) eliminates all other methods in one fell swoop For another thing, most of the methods that have actually been proposed for fixing the intended interpretation of our language fall under premise (2) rather than premise (1)[9] Hence, (2) is the premise which warrants the most sustained philosophical attention

The heart of Putnam's defense of premise (2) is the observation that the phrase 'theoretical constraints' is broad enough to encompass *philosophy* as well as mathematics and natural science In particular, any philosophical account of the way set theory gets its "intended interpretation" can itself be viewed as just one more theoretical constraint Hence, no such account allows us to evade the problems introduced by premise (1) since Putnam can always find an assortment of models that satisfy *both* our original theoretical constraints *and* our new philosophical semantics, the philosophical semantics do not force our language to take a unique "intended interpretation."[10]

---

[8] Note, for instance, that we have to hope that there are no large cardinals stronger than the ones explicitly mentioned in our "theoretical constraints " Similarly, we have to hope that a great many small models (including those witnessing the consistency of our *own axioms*!) do not exist While this is certainly a *consistent* hope, that is about all that can be said for it

[9] For example, causation for language referring to material objects and set-theoretic "perception" for language referring to sets

In effect, then, Putnam argues that premise (2) is *vacuously* true It is not that there are a variety of genuine alternatives to using theoretical and operational constraints to fix the intended interpretation of set theory (and that Putnam's argument for premise (2) takes care of these genuine alternatives) Rather, it is that every *seeming* alternative turns out, upon reflection, to be a *special case* of "theoretical and operational constraints " Hence, every alternative is really covered by the argument for premise (1) In short because the phrase 'theoretical and operational constraints' is flexible enough to *commandeer* rival mechanisms for fixing intended interpretations, all such mechanisms are ultimately taken care of by premise (1)

Now, this argument for premise (2)—the so-called "just more theory" argument—has been extensively discussed in the literature. It is useful, therefore, to outline a standard objection to this argument and to sketch Putnam's response to that objection.[11] The objection comes in two parts First, we distinguish between *describing* the features of a model which make it an "intended interpretation" and simply *adding new sentences* for a model to satisfy in order to count as an intended interpretation More perspicuously, we distinguish between *changing* the semantics under which a collection of axioms (or "theoretical constraints") gets interpreted—for example, by restricting the class of structures that count as "models" for our language and/or strengthening the notion of "satisfaction" that ties sentences to models—and simply *adding new axioms* to be interpreted under the same old semantics

---

[10] This way of formulating the argument may be slightly misleading Putnam thinks that, to the extent that our philosophical semantics are any good, they should *already count* as part of our "original theoretical constraints " Hence, they should *already be covered* by his argument for premise 1

[11] The objection at issue has been raised in various forms by various people, see, for example, Michael Devitt, *Realism and Truth* (Princeton University Press, 1984), chapter 11, David Lewis, "Putnam's Paradox," *Australasian Journal of Philosophy*, LXII (1984) 221-36, and James Van Cleve, "Semantic Supervenience and Referential Indeterminacy," this JOURNAL, LXXXIX, 7 (July 1992) 344-61 Putnam's response to this objection comes in the introduction to his *Realism and Reason*, pp x-xii

It is worth noting that several recent papers claim that this objection rests on a misinterpretation of Putnam's original argument—see David Leech Anderson, "What Is the Model-theoretic Argument," this JOURNAL, XCIII, 6 (June 1993) 311-22, and Igor Douven, "Putnam's Model-theoretic Argument Reconstructed," this JOURNAL, XCVI, 9 (September 1999) 479-90 Although I disagree with this claim, it would take us too far afield to discuss the relevant interpretive questions here I discuss these questions in "Revisionary Interpretations of the Model-theoretic Argument" (in preparation)

Second, we note that many natural conditions on the intended interpretation of set theory fall on the description side of this distinction. When a philosopher claims that the intended models of set theory should be transitive, she is *describing* the structures that are to count as models for her axioms, she is not just adding new sentences to be interpreted under Putnam's favorite models. Similarly, when she claims that intended models should satisfy second-order ZFC, she is explaining *which semantics* (and, more specifically, which satisfaction relation) her axioms should be interpreted under, she is not just adding new axioms to be interpreted under a (first-order) semantics of Putnam's choosing.

Now, if all this is correct, then Putnam's "just more theory" argument is in trouble, since it clearly assumes that all conditions on intended interpretations are really of the "adding axioms" variety. How can Putnam respond? Basically, he claims that this objection begs the question against the model-theoretic argument. The argument, after all, involves the question of whether mathematical language has determinate significance, and the realist seems to assume that it has such significance when she uses phrases like 'transitive', or 'complete power-set' to describe her notion of "intended model" (or to explain her preferred semantics for interpreting axioms). To adapt a passage from *Realism and Reason*.[12]

> Here the philosopher is ignoring her own epistemological position. She is philosophizing as if naive realism were true of her (or, equivalently, as if she and she alone were in an *absolute* relation to the world). What *she* calls 'transitivity' really is transitivity, and *of course* there is a fixed, somehow singled-out, correspondence between the word and one definite property in *her* case. Or so she assumes. But how this can be so was just the question at issue. (*op. cit.*, p. xi)

On Putnam's view, then, as long as the determinacy of mathematical language is still at issue, realists cannot—on pain of begging the question—use this language to describe the intended interpretation of their theories and/or to specify the semantics under which their axioms should be interpreted.

To summarize, Putnam's "just more theory" argument for premise (2) comes in two stages. First, Putnam considers realistic theories of interpretation and argues that these theories can be regarded as additional “theoretical constraints.” As such, they do not add anything new to the “theoretical and operational constraints” of premise (1). Second, Putnam considers the claim that realistic theories *describe* the ways interpretations are fixed (or, equivalently, describe the kinds of semantics under which our axioms are to be interpreted). Since this claim assumes that we can use mathematical language to describe, in a determinate fashion, the intended interpretation of set theory, Putnam dismisses it as question begging.

---

[12] This passage originally concerned the causal theory of reference, I have modified it to fit the mathematical case. In particular, I have replaced the term 'causality' with 'transitivity' in two places, and the phrase 'one definite relation' with 'one definite property'. Finally, I have changed the gender of Putnam's philosopher to fit that in the main text.

IV. JUST MORE THEORY II

In this section, I give three responses to Putnam’s “just more theory” argument. First, I show that this argument is incompatible with some of Putnam’s other theoretical commitments, hence, whatever merits it may have in its own right, it is not an argument which is available to Putnam. Second, I show that analogous arguments can be made in situations where these arguments clearly do not work, and I argue that, by parity of reasoning, Putnam’s argument does not work either. Finally, I give a straightforward explanation as to *why* Putnam’s argument does not work: in brief, I show that Putnam’s charge of question begging rests on a (fairly trivial) logical mistake. At the end of the section, I conclude that Putnam’s “just more theory” argument provides no support for premise (2).

To begin, note that Putnam’s charge of question begging in the last section rests upon an assumption: that it is illegitimate to use semantically *in*determinate language to describe the intended interpretation of set theory. After all, the problem with realists’ attempts to use phrases like ‘is transitive’ or ‘satisfies second-order ZFC’ to describe the intended interpretation of set theory is that we cannot show that these phrases have determinate interpretations (and that simply *assuming* they have determinate interpretations would beg the question). For this objection to have force, however, we must first think that there is something wrong with using semantically indeterminate language to describe the intended interpretation of set theory. Only against the backdrop of this assumption would the realist even *need* to claim that, for example, ‘is transitive’ or ‘satisfies second-order ZFC’ are semantically determinate.

Next, we need three technical facts about first-order logic. First, first-order logic lacks the resources to capture the notions of finitude and recursion. On the finitude side, there is no collection of formulas which characterizes exactly the finite models (or finite subsets of models). On the recursion side, there is no collection of formulas which captures the notion of a *recursive* definition. Second, the notion of finitude is needed to characterize the syntax of first-order logic: the sentences of first-order logic can be of arbitrary *finite* length, but they cannot be infinite[13] Third, the definition of satisfaction for first-order logic is recursive it starts with a definition for atomic formulas and then supplies recursion clauses—one for each connective and quantifier in our language—to extend this definition to arbitrary formulas

Combining these technical facts with the assumption from the penultimate paragraph, we can see that Putnam has got himself into hot water The following four claims are on the table

(1) The notions of finitude and recursion are needed to describe first-order model theory
(2) First-order model theory cannot capture the notions of finitude and recursion
(3) It is illegitimate to use semantically indeterminate notions to describe "intended interpretations "
(4) Only those notions which can be captured by first-order model theory are semantically determinate

Here, claims (1) and (2) are repetitions of the technical facts discussed in the last paragraph, and claim (3) is the assumption on which Putnam's charge of "begging the question" was seen to rest Claim (4) is a version of Putnam's own premise (2) it ensures that the language of set theory is no more determinate than the argument for premise (1) would allow[14]

Together, these four claims cause serious problems for Putnam From claims (3) and (4), it follows that only notions that can be captured by first-order model theory can be used to describe the "intended interpretation" of set theory Combining this with claim (2), we get that the notions of finitude and recursion cannot be used to describe the intended interpretation of set theory But this fact, together with claim (1), entails that first-order model theory cannot be used to describe the intended interpretation of set theory Since this is precisely what Putnam's own arguments require us to do (for example, in the argument for premise (1)), Putnam's position is internally inconsistent

---

[13] Note that this convention is crucial to the formulation of first-order ZFC Two of the axioms of ZFC—comprehension and replacement—are really *axiom schemas* That is, they are really infinite *lists* of axioms, one for every formula in the language of set theory Hence, a firm grasp on the notion "arbitrary formula" is required to *formulate* the axioms of first-order ZFC

[14] That is, it ensures that notions which are not fixed by (the first-order models of) our "theoretical and operational constraints" are not fixed by other means either These notions remain indeterminate

This, then, is what I like to call the *stability objection* to Putnam's argument. The argument rests on the assumption that we cannot use semantically indeterminate language to describe "intended interpretations." But, by Putnam's own standards, the notions needed to formulate first-order model theory turn out to be semantically indeterminate. So, since Putnam's own techniques for obtaining intended interpretations involve first-order model theory, his overall position is logically unstable.

Leaving this stability objection aside, I note that Putnam's argument is also rendered suspect by the company it keeps. Consider the old, childhood puzzle about the three men and the thirty dollars.

> Three men walk into a bar. They order drinks. At the end of the evening, the bartender presents them with a bill for $30.00. Each man pays $10.00, and the three men leave the bar. Once they have left, the bartender realizes that the bill should only have been $25.00, so he gives the busboy $5.00 to take to the three men. The busboy, being a dishonest chap, gives each of the men $1.00 and pockets the remaining $2.00 for himself. At the end of the day, then, each of the three men has paid $10.00 and gotten $1.00 back. Now, 10 − 1 = 9, and 3 × 9 = 27. If we add in the $2.00 in the busboy's pocket, we are left with $29.00. So, what happened to the missing $1.00?

Now, suppose that someone presents this puzzle as a serious challenge to arithmetic—say, as a challenge to the determinacy of ordinary talk about the natural numbers. Suppose also that we respond to this suggestion by giving the obvious solution to the puzzle—that is, by explaining where the puzzle's arithmetic goes awry.[15] Why, in this case, cannot the proponent of the puzzle simply follow Putnam and claim that this solution "begs the question"? After all, the puzzle is supposed to challenge the determinacy of ordinary talk about the natural numbers, and we have to *use* such talk to explain why the puzzle is mistaken.[16]

This, then, gives us an argument that is structurally analogous to (one part of) Putnam's "just more theory" argument. Clearly, however, this new argument does not work. The "paradox of the three men" is not a real paradox, the solution given in footnote 15 is a good solution, the fact that this solution uses number theory is irrelevant, and we do not beg any questions in saying all of this Given the structural similarities between this failed argument and Putnam's own "just more theory" argument, therefore, I think that considerations of parity provide us with ample grounds for rejecting the latter.

---

[15] For those who have not seen this solution, it is really quite simple. The three men *did* pay $27.00. This $27.00 can be split into two parts: $25.00 which is inside with the bartender and $2.00 which is in the busboy's pocket. There is also $3.00 which the men got back from the bartender (via the busboy). As expected, 27 + 3 = 30. The puzzle turns, therefore, on a trick. Instead of adding the $3.00 which the men received to the $27.00 which they paid, the puzzle tries to convince you to add the $2.00 in the busboy's pocket to the $27.00 and to simply ignore the $3.00. Since the $2.00 is *already part* of the $27.00, this is clearly illegitimate.

[16] That is, we have to refer to numbers like 2, 3, 25, 27, 29 and 30. Indeed, we even employ higher-order operations on these numbers: addition, subtraction, multiplication, and so on.

Of course, it would be nice if we did not have to rely on parity—or on theoretical conflicts with *Putnam's* other views—to get a response to the "just more theory" argument Fortunately, it is relatively easy to explain *why* Putnam's charge of question begging misses the mark. In evaluating his model-theoretic argument, we are interested in the question. Does Putnam's model theory entail that mathematical language is semantically indeterminate? To answer this question, we need to evaluate the following conditional.

Putnam's Model Theory $\Rightarrow$ Semantic Indeterminacy

Now, as a general rule, conditionals of the form $P \Rightarrow \neg Q$ are evaluated by asking whether we can accept both P and Q. If we can, then the conditional has to be rejected

In the case at hand, therefore, we ask whether we can accept both Putnam's model theory and the semantic determinacy of mathematical language. That is, we (tentatively and hypothetically) accept both of these things, and we then check to see whether the resulting combination leads to a contradiction In this context, our hypothetical acceptance of semantic determinacy does not constitute an instance of "begging the question." It is just part of the standard method for evaluating ordinary conditionals

Two comments are in order here First, this "evaluation of conditionals" point does not depend on controversial assumptions about the status of the larger dialectic between Putnam and the realist. It does not presume that we have an adequate account of mathematical reference; nor does it rely on fancy considerations regarding the burdens of proof in realist/antirealist debates Indeed, it is compatible with the assumption that mathematical language *really is* indeterminate—that is, that Putnam is *right* on the deep philosophical issue The point simply shows that one aspect of Putnam's model-theoretic argument—namely, his charge of "question begging" in his defense of premise (2)—rests on a misapplication of the logic of conditionals

Second, this "evaluation-of-conditionals" point is completely uncontroversial when questions other than semantic determinacy are at issue Suppose that two philosophers—John and Alvin—are arguing about the existence of God John presents a version of the problem of evil, and Alvin responds by sketching a hypothetical situation in which God co-exists with all of John's evils In this case, Alvin has not "begged the question" against John Even though the overall debate is *about* the existence of God, it is perfectly legitimate for Alvin to evaluate John's argument by constructing this kind of hypothetical Although the hypothetical cannot show that God exists, it can show that John's *argument* fails

So too, then, in the case of Putnam's model-theoretic argument Even though the overall debate between Putnam and the realist concerns the determinacy of mathematical language, it is perfectly legitimate for the realist to assume such determinacy in the course of evaluating Putnam's argument Just as Alvin's argument begs no questions against John, so the realist's argument begs no questions against Putnam To think otherwise is to misunderstand the logic of ordinary conditionals

To sum up, Putnam's "just more theory" argument fails There is a basic difference between *describing* the features that make a model "intended" and simply *adding* new sentences for a model to satisfy Put otherwise, there is a difference between *changing* the semantics under which a set of axioms is interpreted and adding new sentences to be interpreted under the same *old* semantics Further, Putnam's charge of "question begging" does not allow him to undercut—or to slur over—this distinction At best, the charge rest on a misapplication of the logic of ordinary conditionals At worst, it introduces assumptions that get Putnam into even more trouble than he was in before—for example, the stability objection

V SOME CONNECTIONS

In this section, I highlight a connection between the mathematical problems discussed in section II and the philosophical problems discussed in section IV In particular, I show that both kinds of problems involve a certain dialectical assumption that Putnam tends to make, and I argue that realists have no reason to grant Putnam this assumption In light of this, I conclude, once again, that there is little reason for realists to worry about Putnam's model-theoretic argument

Let us begin with a passage originally discussed in section III Putnam is considering the claim that realists use mathematical language to *describe* the intended interpretation of set theory He writes "Here the [realist] is ignoring his own epistemological position He is philosophizing as if naive realism were true of him as if he and he alone were in an *absolute* relation to the world."[17] Now, leave aside the question of whether there is anything *wrong* with such naive philosophizing—a question answered in the negative in the last section—and ask instead whether Putnam engages in any naive philosophizing of his own.

To see the worry here, recall the overall structure of Putnam's argument. He begins by considering a theory that includes "all theoretical and operational constraints," a theory that includes *everything* we might ever need to say and (even) "all true sentences" (MR, 18). Once this theory is in hand, Putnam steps back to examine this theory's semantics: he shows that the theory has many different (first-order) models, and he argues that these models all constitute "intended interpretations" of our language.

It is here, in this turn to semantics, that Putnam's own "naive realism" comes into play. Whatever problems our theory is supposed to have (for example, semantic indeterminacy) these problems vanish when he engages in semantics. Putnam allows himself to refer, naively and absolutely, both to a particular theory and to a whole menagerie of models for this theory. Further, for each of these models, he spells out the relationship between model and theory in a thoroughly realistic manner. In short, he philosophizes as though he and he alone were in an *absolute* relation to the world of models—what *he* calls "models" really are models, what *he* calls "satisfaction" really is satisfaction, and of course there is a fixed, somehow singled-out, correspondence between the language of model theory and the world of models in *his* case.

As we saw in section III, however, this is precisely the standpoint that Putnam denies to the realist. When the realist tries to "stand back" from his set theory to talk about that theory's interpretation—to specify, for instance, that this interpretation must be transitive, or well founded, or satisfy second-order ZFC—Putnam accuses him of "begging the question." Although Putnam's own model-theoretic talk should be viewed as talk *about* set theory, the realist's talk must be viewed as talk *within* set theory.[18]

---

[17] *Realism and Reason*, p. xi.

[18] Clearly, this point amounts to a less-technical reworking of some of the ideas behind the "stability objection" in section IV. Putnam needs to use notions like "finite" to pick out his own preferred "intended interpretations" of set theory; however, these are precisely the notions he wants to deny the realist. Hence, it is only by assuming a stark asymmetry between his own position and that of his opponents that Putnam can make his overall argument hold together. (A note: if Putnam *were* to allow realists to use the notion "finite," then they could define the

This asymmetry between Putnam's own position and the position he allows his opponents resembles closely the asymmetry discussed at the end of section II.1. There, Putnam wanted to limit his opponents to a specific collection of set-theoretic axioms, while he himself used stronger axioms to prove his key theorem Here, Putnam wants to limit his opponents to working *within* a particular theory, while he himself steps outside this theory to talk about its semantics. In both cases, then, Putnam's arguments depend on allowing himself just a little more than he allows those against whom he is arguing.

Clearly, however, there is no reason for realists to go along with all this. As long as realists insist on using the same resources that Putnam himself uses, neither of his key arguments can get off the ground His argument for premise (1) will fail for the reasons discussed in section II, and his argument for premise (2) will fail for the reasons discussed in section IV (and in this section) With both of these arguments so disabled, Putnam's larger model-theoretic argument becomes completely unthreatening.

VI CONCLUSION

I have isolated two flaws in (one version of) Putnam's model-theoretic argument against realism. I have shown that one half of his argument—that is, his argument for premise (1)—rests on a mathematical mistake, and I have argued that he cannot fix this mistake without jeopardizing the very philosophical point that his mathematics is supposed to support. On a more philosophical note, I have argued that the nonmathematical portion of Putnam's argument is just as problematic as the mathematical portion. His defense of premise (2) introduces inconsistencies into his overall position, renders his argument strikingly similar to some obviously faulty arguments, and, in the final analysis, depends on a straightforward misapplication of the logic of ordinary conditionals. What is more, it seems likely that *any* defense of Putnam's premise (2) will face similar problems—for example, will run aground on the "stability objection" of section IV. In the end, therefore, I think that realists have very little to fear from Putnam and his models

TIMOTHY BAYS

University of Notre Dame

---

notion "well-founded " This would be enough to rule out the models generated by theorem 1)

ON SENSE AND REFLEXIVITY

Theories of meaning for names have been stymied by two apparently incompatible constraints. The first is that a theory accept the fact that a name is simply a label for a thing that does not designate it in virtue of properties that the thing exhibits. The second constraint is that a theory explain the fact that replacing a name in a sentence by another name of the same thing yields a sentence that expresses a different proposition from that expressed by the original sentence.

There is a theory of names that succeeds in breaking the impasse created by these constraints. It begins from the proposition that names have word-reflexive meaning and develops the implications of this semantic reflexivity. Each proper name, I shall argue, designates its own bearer. This claim will appear to some to be true but insignificant, while to others it will seem obviously false. It is not false, however, and it has implications for the philosophy of language which are far from trivial. First, an appreciation of the difficulties that face theorists of names will be helpful.

I. IMPASSE

How to satisfy, or to escape, the two constraints has puzzled philosophers since Gottlob Frege. In "On Sense and Reference,"[1] Frege begins from the observation that, if a name had nothing but the object it names as its meaning, then the change of sentence content effected by substitution of codesignative names would be inexplicable. The cognitive value of the sentence 'Phosphorus is Hesperus' clearly differs from that of 'Hesperus is Hesperus'. Substitution of one name of Venus for another produces sentences with different content. Frege's solution is to reject the first constraint.

Frege argues that a name must have, in addition to its referent, a sense. Its sense is a condition whose satisfaction by an object is necessary and sufficient for the name to designate the object. To designate an object by name is to present it as the unique satisfier of the sense condition expressed by the name.

The problem with Frege's solution is that the first constraint seems inescapable. When we look for the senses of names, there seem to be none. Names are simply labels for things. Even Frege admitted that

---

[1] In Peter Geach and Max Black, eds., *Translations from the Philosophical Writings of Gottlob Frege*, 3rd edition (Totowa, NJ: Rowman and Littlefield, 1980), pp. 56-78, here p. 56.

0022-362X/01/9807/351-64

K.f - 1116

for an ordinary proper name like 'Aristotle' there is nothing that seems to be the sense of the name (*ibid.*, p. 58 note). At best, there are various senses that can be attached to the name depending on who is using it. This threatens semantic chaos in both direct and indirect discourse. If the criterion for designation by a name varies from person to person, there is no plausible expectation that the referent will always be the same. Different things may satisfy different criteria. In indirect discourse, propositional attitudes cannot be reliably ascribed if the sense of a name in the mouth of the reporter cannot be expected to match either the sense attached by the one to whom the attitude is ascribed or the senses attached by the hearers of the report.

The view that names do not have senses has been reinforced by Saul Kripke.[2] He argues cogently that names function in natural language in the way that rigid designators function in modal logic. A rigid designator is a term that designates the same individual with respect to every possible world in which that individual exists. For the name 'Aristotle', being a rigid designator means that its designation of Aristotle cannot depend on his possessing properties that he lacks in some possible worlds. This name refers to Aristotle even when it is used to speak of counterfactual circumstances in which he exhibits different properties and stands in other relations.

Given that 'Aristotle' is a natural-language rigid designator, it must not express a sense. A sense condition, it seems, would limit—in a way that should be recognizable by anyone who understood the name—the circumstances at which the name would designate Aristotle. For example, if 'Aristotle' were to have the sense of 'the last great philosopher of antiquity', then, since no one is necessarily a great philosopher, 'Aristotle' would designate someone else at some counterfactual circumstances. Since its reference does not shift with circumstances, it seems that the name does not have a sense.

'Aristotle' would have a sense only if it expressed a generally recognized singular condition such that only Aristotle could satisfy it and such that Aristotle could not exist and fail to satisfy it. The prospect for a sense of an ordinary proper name that meets this standard, which is set by the conception of names as rigid designators, seems dim.

A few years after arguing in *Naming and Necessity* that modal considerations make it implausible that names have senses, Kripke presented a puzzle about names and cognitive content that apparently cannot be solved by positing senses for names[3] On Kripke's view, senses are impotent when faced with a puzzle of the sort for which they were introduced There are two variants of the puzzle

---

[2] *Naming and Necessity* (Cambridge: Harvard, 1980).

First, Pierre, having arrived at a limited set of beliefs about a city both by hearsay in France and by observation in England, comes to utter with normal comprehension and full sincerity both '*Londres* is pretty' and 'London is not pretty' Pierre's problem, as Kripke presents it, is that he uses '*Londres*' and 'London' without knowing that these are the French and English versions of one name The puzzle in this situation is to state the proposition that Pierre believes If 'London' and '*Londres*' are versions of the same name, there are not two senses to distinguish two propositions So, does Pierre believe that London is pretty? The question seems to be unanswerable without unfairly charging Pierre with logical inconsistency

Second, Peter, who utters 'Paderewski had musical talent' with sincerity and comprehension, hears the same name, 'Paderewski', used in reference to a politician He mistakenly believes that the politician is not the musician whom he admires As an expression of his view of the artistic abilities of politicians, Peter sincerely utters 'Paderewski had no musical talent' The puzzle arises again Since there is only one name, even if names had senses there would not be two propositions to be distinct objects of Peter's belief So, does Peter believe that Paderewski had musical talent?

While Frege's puzzle about the cognitive content of identity sentences vanishes if names merely have senses, the solutions to Kripke's puzzles about belief ascription require not only senses of names but also principles of individuation for names

II REFLEXIVE SENSE

The theory of names that solves Kripke's puzzles also satisfies both the constraint that a theory account for a change of proposition with substitution of coreferential names and an appropriately restricted version of the constraint that a theory accept that names are just labels Moreover, it explains how names' rigid designation results from their generally recognized senses

To give the sense of one name is to provide a schema that will serve for any name 'Aristotle' means the same as 'the individual named "Aristotle"' A name presents its referent as the bearer of the name A name plays an essential role in its own condition of satisfaction, which is to say that it has word-reflexive sense

---

[3] "A Puzzle about Belief," in Nathan Salmon and Scott Soames, eds , *Propositions and Attitudes* (New York Oxford, 1988), pp 102-48

This word-reflexive analysis bears a family resemblance to other recent theories that in one way or another give names a place in their own semantical analyses. Kent Bach's[4] *nominal description theory* (NDT) and Jerrold Katz's[5] *pure metalinguistic description theory* (PMT) are perhaps the most notable. According to Bach, a name '*N*' is semantically equivalent to the description 'the bearer of "*N*"'. For Katz, the sense of a proper noun '*N*' has the form 'the thing which is a bearer of "*N*"'

What the members of this family of analyses have in common—that a name designates one who bears it—may make them seem trivial. This, though, is their common strength. If names have senses, they must be obvious to the ordinary user of the language. The Fregean notion of sense is that which is grasped by one who understands the word Names could hardly have senses discernible only by philosophers of language.

The crucial differences within this family of theories are, for the most part, not visible in their descriptive paraphrases of names The mere reference to the name in each analysis conceals a fundamental disagreement about the nature and individuation of names The present word-reflexive analysis will be seen to differ from both Bach's NDT and Katz's PMT on the crucial matter of the nature of a proper name; consequently, it alone will be seen to account for the relevant semantic data simply and completely

III OBJECTIONS

While it may seem obvious to the average person that a name picks out its bearer, most philosophers of language do not accept a word-reflexive account of the senses of names. In some cases, this may be a matter of not seeing what is hidden in plain sight—the purloined sense But this type of analysis of the meanings of names has not been only recently proposed Kripke[6] notes that William Kneale, Bertrand Russell, and Alonzo Church each suggested something of the sort

Philosophers' skepticism about word-reflexive analyses is both general and specific. In general, the argument that names are rigid designators dims the prospect for any theory of senses for names. No account of names' having senses is going to be, or should be, accepted until it accounts for the "rigidity datum " But there are also doubts directed specifically at analyses that give a name a role in its own sense It is in response to these specific objections that the important differences between the word-reflexive theory advocated here and its relatives, NDT and PMT, will first appear

---

[4] *Thought and Reference* (New York Oxford, 1987), p 135

[5] "Names without Bearers," *The Philosophical Review*, CIII, 1 (January 1994) 1-39

[6] *Naming and Necessity*, pp 68-70, "A Puzzle about Belief," p 139 note 12

In *Naming and Necessity,* Kripke rejects a word-reflexive analysis proffered by Kneale.[7] Denying that ordinary names are signs without sense, Kneale said that 'Socrates' means 'the individual called Socrates' (*ibid.*, p 630). Kripke argues that this account violates the "noncircularity condition" that a name's sense not "involve the notion of reference in a way that it is ultimately impossible to eliminate " To Kripke, Kneale's claim is that when one uses 'Socrates', "he refers to the man to whom he refers."[8]

Kripke's objection assumes one of two possible readings of the participle 'called' in Kneale's sentence. It may be read, with Kripke, as expressing a three-place relation among a speaker, a term, and the speaker's intended referent; or it may be read as 'named' (a two-place relation between a name and its bearer). Kneale's analysis would be misguided if it were that 'Socrates' means 'whichever individual the speaker calls "Socrates"'. This account would give the speaker's reference rather than the meaning of a name.[9] If Kneale's brief remark is to be interpreted charitably, it must be taken to be that 'Socrates' means 'the bearer of "Socrates"'. This interpretation has the additional advantage of allowing Kneale's omission of quotation marks around 'Socrates' in his statement of the name's meaning to be seen as deliberate and careful. To say that a person is called Socrates (using the name rather than mentioning it) is to say that he is Socrates—the bearer of 'Socrates'. He is Socrates by name In contrast, to say that someone calls a person 'Socrates', or refers to him by 'Socrates', is to leave open the possibility that the name is misused He may not be Socrates

The word-reflexive sense is the epistemic basis for determining the referent. The referent is to be found, not by looking at language-independent properties of individuals, but by looking for the individual on which the name was bestowed. A name is a word that has been created, usually by a deliberate performative nomination,[10] to be a context-independent means of picking out some thing or person David Kaplan[11] argues in his Ryle lectures that each proper name is an artifact created, transmitted, and stored by language users for the purpose of speaking to or about one particular individual It is not an especially challenging task, in normal cases, for us to keep track of the individual that bears a particular name

---

[7] "Modality De Dicto and De Re," in Ernest Nagel, Patrick Suppes, and Alfred Tarski, eds , *Logic, Methodology, and Philosophy of Science* (Stanford University Press, 1962), pp 622-33

[8] *Naming and Necessity,* pp. 68-70

[9] Bach also notes that Kripke "conflates the meaning of a name with its use to refer" (*op cit*, p 160)

[10] "Performative nomination" is taken from John Lyons, *Semantics,* Volume 1 (New York· Cambridge, 1977), p 217 Deliberate nomination is not the only way to create a name Many people bear a name that has its origin in a mistaken use of a name that is not theirs Such a mistake can be wittingly or unwittingly "co-opted," that is, reacted to as if it were the intentional introduction of a new name

Keeping track of which individual bears a name is like keeping track of which day bears a date There is nothing about a day, other than its relation to a calendar, that makes it the bearer of a date Ordinarily, repeated and shared use of a calendar in the designation of days suffices for keeping track of which day bears which date It is possible to lose track, but it seldom happens [12]

Kripke came close to the truth about names when he said, in concluding his criticism of Kneale's view, that "whatever this relation of *calling* is is really what determines the reference and not any description like 'the man called "Socrates"' "[13] Indeed, the relation of naming (to avoid the ambiguity of 'calling') is really what determines the reference, but this is the relation that is expressed both by the name 'Socrates' and by those descriptions which mention the name in an expression of its word-reflexive sense for example, 'the bearer of "Socrates"'

The other recent theorists who give a name a place in the analysis of its meaning cannot concur in this straightforward response to doubts about whether word-reflexive sense satisfactorily determines reference The problem is that both Bach and Katz believe that proper names can, and often do, have multiple bearers This means that, according to them, the condition of being a name's bearer cannot be relied on to distinguish a referent While both think that the bearing of a particular name often fails to distinguish an individual, they disagree about why names can have multiple bearers and about how singular reference is achieved when a shared name is used

Katz holds that names are general terms whose tokens have singular reference He uses the unexpected indefinite article in his definite description, 'the thing which is a bearer of "*N*"', because he holds

---

'Madagascar', the name of the island off the coast of Africa, is said to have an inadvertent origin arising out of a mistake made by Marco Polo See Evans, "The Causal Theory of Names," *Aristotelian Society Supplementary Volume,* XLVII (1973) 187-208

[11] "Words," *Aristotelian Society Supplementary Volume,* LXIV (1990) 93-119

[12] It happened to Robinson Crusoe It is hard to keep track of the bearers of dates and names all by yourself

[13] *Naming and Necessity,* p 70

that, semantically, proper nouns (names) are general terms that are distinguished from common nouns by having metalinguistic senses (*op. cit.*, pp. 5-6). For Katz, it is only a use of a name that is singular and that can be characterized as having the form represented by his definite description. The name with its bearer condition serves merely as a first filter restricting the possible referents of any use of the name to the bearers of the name (*op. cit.*, p. 21). The speaker must rely on extralinguistic information, and perhaps even additional description, to make clear which one of the name's bearers is the intended referent of his use of the name.

Bach, by contrast, does not deny that a name is a singular term. A name is "semantically equivalent to" a definite description of the form 'the bearer of "N"'. Such descriptions can determine referents. He believes, however, that there is a problem that frequently makes them inadequate for determining a referent. A name can be given more than one bearer. When a name gets multiple bearers, Bach believes that it becomes semantically equivalent to an incomplete definite description (*op. cit.*, p. 136). An incomplete description is one like 'the table' that, because there are many tables, does not describe a unique object. Bach holds that when incomplete descriptions or names with multiple bearers occur, they are used nonliterally. He sees two ways to use names nonliterally. There is attributive nonliteral use, in which case the name is "used as short for (not synonymous with) a more elaborate expression" which is some complete, or proper, description of an individual (*op. cit.*, p. 136, cf. p. 105). The alternative, and more common, nonliteral use is referential. In this case the semantic inadequacy of a name with multiple bearers is compensated in just the way that Katz thinks a general term is given a singular use. The speaker uses the name while having "some further information in mind on which his audience is to rely, information which is not part of the meaning of the name" (*op. cit.*, p. 170, cf. pp. 105-06).

Some version of Bach's or Katz's pragmatic supplementation to the semantics of names is required only if phonetic and orthographic forms (that is, vocables) are sufficient to individuate names. But there is remarkably little justification for counting names by counting phonetic forms, while there is good reason to individuate them in the way in which words are normally distinguished—even when they share vocables.

The chief reason for being tempted to individuate names merely by vocables is that this way of counting names appears to be implied by the fact that many individuals have the same name. But the adjective 'same' is notoriously elliptical in meaning. Assertions of sameness may be either attributions of numerical identity or attributions of qualitative or generic identity. If a son has his father's nose, no one thinks that there is only one nose between them. They have the same kind of nose Their noses are numerically distinct. No one fails to distinguish between the kind and the individual where the kind is abstract and the individuals are concrete. Where, as in the case of names, both the kinds and the individuals are abstract, it can be easier to mistake identity of kind for numerical identity. Yet the kind-individual distinction holds even for abstract entities When a son has his father's name, their names are of the same phonetic and orthographic kinds; but the father's name was bestowed (created) before the son's, has an altogether different history of occurrences, and has a different bearer. Despite their identity of form, the names will seldom be confused with each other In directories, they will be two names

Mere vocables are not words. They are entities abstracted from words. To linguists, it is an "uncontroversial premise" that a word is at least a triple—phonology, syntactic category, and meaning [14] In modern languages, a word also has a written form. Distinct words can share pronunciation, spelling, and syntactic category. Equiform words with different origins and meanings are homonyms Names originating on different occasions and having their currency preserved through different chains of occurrences are distinct names that necessarily have distinct word-reflexive meanings. The mere fact that other individuals come to have names of the same form does nothing to change a proper name or its reflexive sense The only thing changed by the existence of equiform names is the presemantic task of word identification [15]

Like other individuals, names can be identified in various ways What suffices for the identification of a name when its form is not unique will vary with the situations in which identification is needed A name is frequently identified simply as the one having a particular previous occurrence. Often a name can be distinguished from equiform names by identifying its bearer What is said when a name is used, together with the form of the name, is often sufficient to identify the bearer. On some occasions, a name will even be recognized as the one introduced in a particular act of naming that is either recent or widely known Different names have different origins, different histories of occurrence, and usually different bearers. Each of these features is a source of hints and clues for identifying a name where other names of its form might have occurred. We ordinarily cope rather easily with homonymy in names

---

[14] Pauline Jacobson, "The Syntax/Semantics Interface in Categorial Grammar," in Shalom Lappin, ed , *The Handbook of Contemporary Semantic Theory* (Cambridge Blackwell, 1996), pp 89-116, here p 90

[15] Kaplan, "Demonstratives," in Joseph Almog, John Perry, and Howard Wettstein, eds , *Themes from Kaplan* (New York Oxford, 1989), pp 481-563, here p 559

What individuals share is not a proper name; rather, it is a kind or genus of name.[16] Following Kaplan,[17] these name kinds might be called "generic names " A name's form (the generic name) may be chosen to honor another individual with a name of that form, or it may be chosen to evoke an attractive quality expressed by an adjective or an abstract noun of that form, but a proper name, with its own origin, bearer, and subsequent history, will be identical neither with any other proper name nor with any adjective or abstract noun.

The difference between names and name kinds is reflected in syntax A name does not take determiners and plural endings the way a common noun does [18] When we want a general term that will apply to those whose names are equiform, we "recategorize" a name to make the noun that we need [19] (Do you know a Chloe?)

The fact that names are proper to their bearers allows us to avoid a dilemma offered by both Bach (*op cit*, p. 137) and Katz (*op. cit*, p. 18): that either names are univocal, sharable terms made singular in use, or they must be ambiguous in as many ways as they have bearers. To avoid this dilemma, three ways of construing names must be distinguished First, there is Bach's and Katz's idea that a name can be common to multiple individuals but made proper in use Second, there is the notion, which Bach and Katz rightly regard as absurd that multiple individuals share a single polysemous name—with no plausible account of all these different meanings. Third, there is the view advocated here: that, while names are seldom unique in phonetic or orthographic form, each name has at most one bearer.

Recognition of the distinction between a name and a kind of name leads to a simple, intuitively correct solution for the "Peter" variant of Kripke's puzzle What Peter did not know was that in the political conversation he heard the very name, 'Paderewski', that he knew to be borne by a talented musician. He thought that he was hearing a different name—one homonymous to the musician's name Under this misapprehension, he uttered 'Paderewski had no musical talent'[20] Sincere utterance of a sentence will yield an utterer's belief by disquotation only when there is no linguistic confusion causing him to say something that he does not believe[21] Since Peter had misidentified the name that he used, he did not know what he was saying

---

[16] The family's name is a shared element in the "full name" of a person The use of only the family name to refer to a member of the family is a common form of synecdoche

[17] "Afterthoughts," in *Themes from Kaplan*, p 574, "Words," p 111

[18] Steven Pinker, *Words and Rules* (New York Basic, 1999), p 164

[19] Christopher Lyons, *Definiteness* (New York Cambridge, 1999), p 21

IV EXPLANATORY SCOPE

Thus far, the view that names have word-reflexive senses has been shown to be not circular, not epistemically vacuous, and not short of the requisite number of names Now, does it have the explanatory capacity needed in a theory of names? It solves Frege's puzzle about identity statements, the "Pierre" version of Kripke's puzzle about belief ascription, and the puzzle about how bearerless names can be meaningful Best of all, it explains why names are rigid designators

The content of the discovery that Hesperus is Phosphorus, on the word-reflexive account, is that the object known as the bearer of 'Hesperus' is identical with the object named 'Phosphorus' Two modes of presentation present one thing The proposition that Hesperus is Phosphorus is distinguished from the proposition that Hesperus is Hesperus Even when names are the terms in an identity statement, Frege's puzzle about its cognitive value is solved in Frege's way by appeal to the distinct senses of the terms

Moreover, this identity statement is planetary—not metalinguistic Unlike the *Begriffsschrift* analysis that Frege repudiated, the names are not "themselves under discussion" (*op cit*, p 56) It is the planet that is designated by terms that express two distinguishing attributes bearing the name 'Hesperus' and bearing the name 'Phosphorus' In general, bearing some name is not a negligible property of a thing Imagine the plight of an astronomer of today who, because of unusual isolation, has not discovered one crucial property of an object that he studies its bearing the name 'Venus' Until he discovers that his object has this "handle," he will be doing astronomy under a severe handicap

The word-reflexive analysis provides a similar solution to the "Pierre" puzzle Pierre, unlike Peter, has not mistakenly thought that there were two names where there was only one He has learned two names—not two versions of one name Names have to be individuated as words are ordinarily individuated '*Londres*' and 'London' are two words one belongs to the French language and the other to

---

[20] Kaplan provides the same diagnosis of Peter's confusion—"Words," p 108

[21] "A Puzzle about Belief," pp 130-31

English They have different phonetic and orthographic forms They were introduced into their respective languages on different occasions Consequently, Pierre is able to believe that *Londres* is pretty and that London is not—just as the Babylonian astronomers had two names that enabled them to attribute, without logical inconsistency, incompatible properties to Venus

Kripke takes '*Londres*' to be a translation of 'London' But the fact that names have word-reflexive senses renders them incapable of translation by other names, which will have their own word-reflexive senses Kripke alleges that it is "desperate" to deny that '*Londres*' is the French translation of 'London', but such pairs of names appear to be translations of one another only because they are so frequently substitutable without disturbing truth value, and even when they are not, the substitution is almost always helpful rather than misleading Substitution of names with identical bearers preserves truth in extensional and intensional (modal) contexts It is only in propositional attitude contexts that problems can arise Even in these "hyperintensional" contexts, when the subject of the propositional attitude, unlike Pierre, knows that the two names have the same bearer, substitution will be truth preserving Where the subject of the attitude is monolingual, the use of a name from another language in an attitude ascription will make the ascription literally false, but even so, those hearers or readers who know only the substituted name are provided with a referentially and structurally parallel proposition that seldom misleads No Babylonian ever believed precisely that Phosphorus and Hesperus are distinct heavenly bodies, but presumably some of them had a belief about Venus that was structurally parallel to this The exact content of their mistaken belief involves Babylonian names that are of no concern to most of us [22]

The theory that names have word-reflexive senses also explains why sentences containing bearerless names are meaningful Names are created in acts of naming, some of which fail to give the new name a bearer One who mistakenly believes in the existence of something, or who wishes to deceive others into a false existential belief, may introduce a name without a bearer This name has its word-reflexive sense and can occur in meaningful sentences Lacking a bearer, it can even occur as the subject term in true negative existential sentences. Suppose that a thief, inventing a fictitious provenance for his loot, tells a detective, 'Jim Jones gave it to me'. The detective can truly reply, 'Jim Jones does not exist'. On the word-reflexive analysis, this is to say that there is no one named 'Jim Jones'. Names do not need bearers in order to be meaningful.

---

[22] While it is obvious that Peter's is a one-name problem and that Pierre's is a two-name problem, there will be cases where it is not clear how many names are involved Do French speakers and English speakers have different names for the capital of France, or do they employ variant pronunciations of one name? There is no reason to think that a sharp line can always be drawn, or needs to be drawn, between variants of one name and different names The Peter and Pierre sorts of cases blur into each other

Here, again, the importance of the distinction between name and kind of name appears. There are many Jim Joneses, but none of them bears the name that the thief introduced when he just "made up" a person. The detective reuses the name created by the thief.

Perhaps the most unexpected and important confirmation of the word-reflexive theory of senses for names, given the pervasive influence of Kripke's modal argument against senses for names, is that it entails that natural-language names are rigid designators. A term is a rigid designator just in case, first, it will not change its referent when a sentence in which it occurs is evaluated at different circumstances and, second, it will not fail to refer when a sentence in which it occurs is evaluated at a circumstance in which the object that is its referent at other circumstances exists.

Word-reflexive sense determines reference by means of a property that cannot be exhibited by different individuals in different circumstances. The word-reflexive sense picks out just the object, if any, on which the name was bestowed: the bearer of the name. A name could have another referent only if it could have another bearer, but a name with any other bearer would be another name with its own origin in the naming of that other bearer. Having the bearer it has is an essential property of a name.

Furthermore, a name will not fail to refer to its bearer when a sentence containing it is evaluated at a circumstance in which that individual exists. The existence of its bearer in a circumstance of evaluation is the sole condition for a name's referring with respect to that circumstance. The name itself does not have to be used in the circumstance of evaluation. Aristotle is so designated in virtue of the two-place relation between him and 'Aristotle'—not some three-place relation among him, his name, and a speaker or collection of speakers. This means that the sentence

(1) Aristotle is not the bearer of 'Aristotle'

is not true and could not have been true—if the name used is the one mentioned. There is no circumstance at which a name designates anything other than its bearer.

In contrast, the sentence:

(2) No one gave Aristotle the name 'Aristotle'

is not true, but it could have been true. (2) expresses a three-place relation among speakers, Aristotle, and a name. A name's bearer—its referent with respect to a circumstance of evaluation—may not go by that name in the circumstance.

That Aristotle is not known as Aristotle in a possible circumstance means only that the name is not used there. Bearing a name is like having a length in meters. They are both relations to conventionally established abstract objects. These abstract objects do not have to be known or used in a counterfactual circumstance in order for us to characterize things in the circumstance in relation to them. Our descriptive resources are not confined to the resources available to speakers in the circumstances of which we speak.

A name is a rigid designator because it designates its bearer—which it has, or lacks, essentially. Speakers may not use a name, may forbid its use, may give its bearer another name, or may forget a name; but a name cannot change or lose its bearer.

Theories such as Bach's NDT and Katz's PMT, holding that names have extensions that vary with possible worlds, must reject Kripke's thesis that names are rigid designators. Bach and Katz explain the apparent rigidity of names as an illusion due to the rigid use of names by speakers. Bach says that it is the speaker's "intention that makes it the case that the name (that is, the speaker's use of it) refers to the same individual in all possible worlds" (*op. cit.*, p. 169). Katz claims that "the rigidity of an utterance .derives from the user's 'stipulation' that he or she is using that token of the proper noun .to speak of the same contextually specified individual in every possible world" (*op. cit.*, pp. 14-15).

It is hard to conceive how, by withdrawing some intention or stipulation, a speaker could use a name nonrigidly. Neither of these theorists provides an example of a proper name used to designate an individual nonrigidly. Generic names can be modified by restrictive clauses to form nonrigid designators. for example, 'the Fido that you saw'; but this is done semantically—not by a change of intention or by stipulation. It seems clear that proper names, which cannot be modified by restrictive clauses, are simply rigid designators. They are not used rigidly and nonrigidly.

Unlike NDT and PMT, the analysis presented here accepts both Kripke's modal doctrine that names are rigid designators and Frege's semantic doctrine that names have senses that determine their reference. It appears that Frege failed to find the senses of names because he was looking for conditions that existed independently of words He once remarked that “the same sense has different expressions in different languages or even in the same language” (*op cit*, p 58) This overlooks the fact that an object may have the attribute of bearing a name This attribute is expressible only by the name or by definite descriptions that mention the name Consequently, the sense of a name will not have an expression in another language, unless the name itself is treated as part of that language or is mentioned in a description otherwise expressed in that language

V CONCLUSION

The sense of ‘Aristotle’ has been found, and it divides the possibilities concerning Aristotle from the impossibilities at just the right place Aristotle might not have been a philosopher He might not have been given the name ‘Aristotle’ He might not have existed What is not possible is that ‘Aristotle’ have a bearer other than Aristotle So it is not possible that Aristotle exist in a circumstance and not be the referent of ‘Aristotle’ when it is evaluated with respect to that circumstance The word-reflexive sense meets the standard set by the conception of names as rigid designators It is a generally recognized singular condition, which can be satisfied only by Aristotle, and it is not possible that Aristotle exist and fail to satisfy the condition

There is a final confirmation of the word-reflexive analysis it satisfies what is correct in the first constraint on a theory of names The first constraint, properly restricted, is that a theory accept that a name is a label for a thing that does not designate it in virtue of properties other than its relation to the name

JOHN JUSTICE

Randolph-Macon Woman's College

COMMENTS AND CRITICISM

ON MULTIPLE REALIZATION AND THE SPECIAL SCIENCES

In "Multiple Realizations," Lawrence A Shapiro[1] aims to protect the special sciences—that is, biology, psychology, the behavioral and the social sciences—from the consequences of his denial that multiple realization is widespread in nature

It is widely held that disciplines are autonomous when their taxonomies are "substrate neutral" and when the events, states, and processes that realize their descriptive vocabulary are heterogeneous This will be particularly true in the case of disciplines whose taxonomies consist largely in terms that individuate by function Having concluded that the multiple realization of functional kinds is far less widespread than assumed or argued for, Shapiro cannot avail himself of the argument for the autonomy of the special sciences which relies on multiple realization This makes urgent the question of whether we must "now give up the idea that functionalist taxonomies have any scientific value" (650) He acknowledges that we must either deny that the special sciences are autonomous, because higher-level kinds have only a single realization and can thus be reduced, or else we must deny that there are empirical laws in the special sciences "In other words, either special sciences have no ontological independence from lower-level sciences or, worse, they have no empirical laws, which is to say that they are not empirical sciences at all" (650) Shapiro's reductionist/eliminativist dilemma for the special sciences is unreal For he has not canvassed the most important source of multiple realization in nature, and this source obviates his dilemma for most of the special sciences Moreover, the route he offers between the horns of his dilemma leads pretty directly to impalement on its eliminativist horn Or so I shall try to show in this comment

I THE SOURCE OF MULTIPLE REALIZATION IN NATURE MAKES FOR CAUSALLY RELEVANT DIFFERENCES

Shapiro argues that, if a disjunction of physical states shares a distinct, common-causal feature in virtue of which each realizes a functional state, then the functional state will be reducible after all, in spite of the multiplicity of its realizers Thus, no direct inference from multiple realization alone to irreducible autonomy can be vouched safe But Shapiro has overlooked the fact that the source of multiple

---

[1] This JOURNAL, XCVII, 12 (December 2000) 635-54

0022-362X/01/9807/365–73

realization in nature has at the same time assured that, in the long run, multiple realizers will have causally relevant differences from one another, or at least causal differences great enough to obstruct the sort of empirical laws that Shapiro requires for an empirical science.

The ubiquity of multiple realization across the domains of all the special sciences is the result of natural selection.[2] Multiple realization kicks in as soon as natural selection begins operating on physical processes. Since humans are biological systems, the domains of all the special sciences that deal with humans are fated to the multiple realizability that the evolutionary origin of homo sapiens reflects.

Natural selection "chooses" variants by some of their effects, those which fortuitously enhance survival and reproduction. Selection for adaptation and function begins at a relatively low level in the organization of matter. As soon as molecules develop the disposition—chemically, thermodynamically, or catalytically—to encourage the production of more tokens of their own kind, natural selection comes into force. Some molecules become replicators—template or catalyze or otherwise encourage the production of copies of themselves, and these molecules interact with their environments so that changes in them—mutations—will result in changes in their rates of replication in their environments As a result of random physical processes—mutations—among such replicating and interacting molecules, there are frequently to be found multiple *physically distinct* structures with some (nearly) identical rates of replication, different combinations of different types of atoms and molecules, which are about equally likely to foster the appearance of more tokens of the types they instantiate. It is the nature of any mechanism that selects for effects that it cannot discriminate between differing structures with identical effects When natural selection encourages variants to become packaged together into larger units, the adaptations become functions.[3] And functional equivalence combined with structural difference will always increase as physical combinations become larger and more physically differentiated. This blindness of selection to structure is no guarantee of causal heterogeneity in the realizers. But it certainly makes it overwhelmingly probable. Over finite periods of time, nature can select at most for functional similarity, not perfect identity. Any two or more physical systems that solve a "design problem" well enough to allow for some minimal level of survival and reproduction rates will be selected. If there are at least two local optima in an adaptive landscape, and if mutation, recombination, drift, and selection can find them, the result will be causally different multiple realizations.

---

[2] Shapiro recognizes that natural selection has some bearing here He writes at one point. "Natural selection has figured out different psychological means to the same psychological ends" (642). But this statement figures as a mere concession to exponents of multiple realization of mental states

[3] I adopt here the "selected effects" account of function due originally to Larry Wright, *Teleological Explanation* (Berkeley California UP, 1976) Shapiro appears to be committed to this account as well.

There is another aspect of natural selection that makes causally heterogeneous realizers highly likely. Evolutionary "design problems" have the reflexive character of what Richard Dawkins[4] and others have called "arms races," dynamic strategic competitions in which every move generates a countermove so that conditions are never constant. From an early point in evolution, the environments that select for any one function began to change more rapidly. For each new solution to a design problem—each new functional trait—becomes part of the environment, setting a new design problem for other functional traits—within the same biological lineage or beyond it in a competing lineage. In the competition for limited resources endemic to the biosphere, any variation in the lineage of a gene, individual, group, or species which enhances fitness (that is, adaptation/function) will be selected for. Any response to such a variation within the heritable repertoire of the competitor gene, individual, group, or species will in turn be selected for by the spread of the first variation, and so on. If the "space" of adaptational "moves" and countermoves is very large, in the long run every functional kind will at least sometimes be realized by multiple causally distinct realizations.

What this means, of course, is that any generalization linking a single functional kind to a set of physical realizations will be at best a temporarily true claim about a historically limited pattern to which, over evolutionary time periods, the number of exceptions will mount until its subject becomes extinct. Take a simple example, such as 'All genes are composed of DNA molecules'. The discovery of the retroviruses showed that this generalization needed to be revised to 'All genes are composed of DNA or RNA molecules'. But RNA and DNA are both composed of the same nucleic acids (qualification: not exactly the same, but similar ones with different effects on information fidelity in the case of uracil and thymine). Presumably, it is this common causal fact about them which accords them the same function of hereditary information storage and developmental control. Alas (even leaving aside the significant qualification), the discovery that prions (proteins which cause BSE or "mad cow disease" and catalytically transmit their (secondary structure) shapes across the generations down a line of descent) shows that the molecules with hereditary function are not all nucleic acids. But, of course, the prion's stratagem for persistence leads to a counterstroke by the nucleic acid-lineage (for example, some nucleic acid based organisms cease feeding on the brains of their victims, or cease feeding their cattle ground-up prion-infected offal). There is a vast space of available adaptive strategies among competing molecules, genes, organism, populations, species. Therefore, fulfilling the same function cannot forever remain a matter of employing the same or similar causal mechanisms. At most one can claim that multiple realizers may only temporarily have a common causal feature which results in their instantiating a functional type. Eventually, mother nature will find a causally new way to perform the function and the class of realizers will no longer be causally homogeneous.

---

[4] *The Blind Watchmaker* (New York: Norton, 1986), p. 178.

Are temporary empirical regularities good enough to allow for the reduction of the functional kinds of the special sciences to their physical realizers Shapiro hopes for? After all, in the context of evolutionary change, 'temporarily' might be a matter of millions of years.[5] Thus, to use Shapiro's example (646-47), over the period that interests physiology, the realizers of visual image-formation may be physically diverse while sharing a single feature (the lens) causally relevant to image formation. But notice several things. First, in the domain of any of the special sciences, where competing systems respond to one another's stratagems more rapidly, natural selection pretty well assures that the realizers of some functional kind will become different in casually relevant respects over time scales that make even "temporarily true" reductions unlikely. The realization bases of the functional kinds of social and behavioral science will certainly not stand still long enough to allow for the casual homogeneity and consequent reduction Shapiro envisions. Second, the "generalization" expressing the reduction of a function will be a temporary, historically bounded one, not a law of the sort Shapiro requires (654). Such a historically limited temporarily true statement will have explanatory power *by itself* only on a novel account of scientific explanation (I return to this point in section II).

Third, and most important, it is unlikely that even temporarily true reductions will be available for the functional kinds that interest

---

[5] This is one upshot of Marc Lange's argument in, "Are There Natural Laws concerning Particular Biological Species?" this JOURNAL, XCII, 8 (August 1995): 430-51.

Shapiro most. For he is particularly eager to undermine the inference from the multiple realization of mental states by brain states to autonomy of psychology from neuroscience. But if the relationship between molecular biology and genetics, development, and cell physiology is any guide, the prospects for causal homogeneity are low. The molecular realization base of classical genetics is already sufficiently causally heterogeneous as a result of natural selection over three billion years that all reductive generalizations linking genetic functions to molecular realizations are riddled with exceptions. As in genetics, reductions in functional anatomy may start out looking smoothly reductive: “genes code for hereditary information because they are composed of nucleic acid bases,” “eyes produce visual images because they are composed in part of (physical) lenses.” But the more detailed and complete the explanation we seek of how a gene, an organelle, a cell, a tissue, or an organ accomplishes its function, the more disjunctive and causally heterogeneous the explanation becomes. Since cognition is much more complex than vision, digestion, fission, meiosis, or protein synthesis, we should expect evolution to have made its realizations more heterogeneous by orders of magnitude. And it is this heterogeneity which reducing the psychological to the neural must deal with. After all, the search for psychophysical laws will not be satisfied by gross anatomical localization of encapsulated cognitive modules. We have this already to some extent. Psychophysical laws need to link occurrent intentional states with their neural realizations. Shapiro asks: “Without considerable knowledge about how the brain produces a mind, why be so sanguine that there are in fact many *truly* distinct ways to build a mind” (646)? The answer to this question may now be a little clearer. In the philosophy of psychology, the multiple realizability thesis is a hypothesis advanced to explain the absence of discoverable psychophysical laws in a way compatible with physicalism. True, the best evidence for it must eventually come from detailed studies of the brain, but initially, we have a good deal of indirect evidence for it from elsewhere in biology.

II. HOW NOT TO PRESERVE THE AUTONOMY OF THE SPECIAL SCIENCES

Shapiro’s dilemma consists in what he calls “two undesirable results” (650):

> Horn 1, *reductionism*: functional, or as he says, “higher-level kinds” have single realizations, ergo no autonomous empirical laws about higher-level kinds, and so no special sciences

versus

> Horn 2, *eliminativism*: higher-level kinds have multiple realizations, thus no empirical laws connecting higher-level kinds with realizers, and a

> fortion no laws connecting higher-level kinds with one another, ergo no empirical laws at all, autonomous or otherwise, and no autonomous special sciences

It is worth noting that horn 1, reductionism, is by no means a clearly "undesirable result." Indeed, Shapiro later seems to admit as much (653). But if the arguments of section i above are correct, in most domains laws expressing the reduction of higher-level kinds of special sciences to lower-level kinds that Shapiro envisions (654) are not on the cards, since the higher-level kinds will be multiply realized by causally heterogeneous mechanism and processes.

One way to circumvent this dilemma is to deny that laws, empirical or theoretical, are indispensable for a science, special or otherwise. This is a very popular "move" in contemporary philosophy of science, especially in the philosophy of biology. In large measure owing to the fact that biology is bereft of laws of the sort we are familiar with from physics, philosophers of biology have been at pains to argue that laws are not required in biology. If such arguments are accepted for biology, they will be equally available for the special sciences [6]

Denying that there are laws in biology is unattractive because it comes with a concomitant obligation to provide a new account of how explanation proceeds in biology in their absence. In light of the fact that the autonomy of the special sciences is more precisely their *explanatory* autonomy, the obligation to provide such an account of explanation becomes quite pressing for those attracted to this option. The account will have to show how an autonomous special science can explain without invoking proprietary laws even implicitly

Shapiro does not opt for this way out of his dilemma, however; rather, he repeatedly insists on the indispensability of laws in a science But his way of attempting to pass between the horns of the dilemma in fact securely impales him on the eliminativist horn. His solution is to argue that functional kinds have important roles in science even when they have no common effects He invites us to consider the functional type 'eye', on the assumption that it is multiply realized by causally different underlying mechanisms

---

[6] Philip Kitcher writes "success in achieving exceptionless generalizations is by no means a sine qua non for good science Darwinian evolutionary theory has served us as an example"—*Advancement of Science* (New York Oxford, 1992), p 121 According to Elliott Sober, "*general* source laws are hard to come by in evolutionary theory"—*The Nature of Selection Evolutionary Theory in Philosophical Focus* (Cambridge MIT, 1984, reprinted, Chicago University Press, 1994), p 51 See also John Beatty, "The Evolutionary Contingency Thesis," in G Wolters and J Lennox, eds , *Concepts, Theories and Rationality in the Biology* (Pittsburgh University Press, 1995), pp 45-81, and Lange, footnote 3

> Despite the fact that we can say nothing more about eyes (in general) than that they have the function to see, it is still a matter of interest what kinds of things can see We derive a deeper understanding of a given kind of eye from its comparison with other organs that are capable of the same function Thus, for instance in trying to understand how a camera eye works it is useful to study its similarities and differences with the compound eye Doing so provides evidence about which of the properties of the camera eye are causally relevant to the function of seeing and which are not. In short, functional kinds in the special sciences, despite the fact that they never enter into (nonanalytic) laws of the form 'All *P*s are also ___', are of value because they collect and order the domain of a special science in a way that facilitates its investigation (654).

In short, the taxonomies of the special sciences describe classes of explananda. And their value consists in this role. But this rationale for the kind terms of the special sciences will not do the work that the exponents of the autonomy of special sciences require. It only forces Shapiro to impale himself on horn 2, eliminativism.

The whole point of the argument from multiple realizability to the autonomy of the special sciences is that the taxonomies of these disciplines include kinds which are *explanatory*, which figure in explanantia, not simply in explananda. Exponents of the autonomy of the special sciences (especially psychology), from Jerry Fodor to the present, have sought explanatory autonomy for these disciplines They have held that we would miss generalizations if we surrendered their taxonomies and that, missing these generalizations, we would be unable to provide adequate explanations for the processes in the domains of these disciplines [7] The point is particularly clear in the philosophy of biology. Consider two examples of how proponents of the autonomy of biology state the view

Elliott Sober[8] writes

> .what answer can we give to the question of whether physics can explain everything that biology can explain [an example of horn 1 of Shapiro's dilemma]? First, we need to divide the question in two (1) If there is a biological explanation for *why some particular event occurred*, is there also a physical explanation? (2) If there is a biological explanation of *what several particular events have in common*, is there also a physical explanation? Perhaps the answer to (1) is yes, as for (2), the answer I would give is no (*ibid.*, p 78)

---

[7] See, for instance, Patricia Kitcher, "In Defense of Intentional Psychology," this JOURNAL, LXXXI, 2 (February 1984) 89-106, which invokes analogies to the explanatory role of multiply realized functional kinds in biology, as illustrated below

[8] *The Philosophy of Biology* (Boulder Westview, 2000, second edition)

According to Sober, biology provides explanations, not just explananda, and these explanations are autonomous because their concepts are 'Fitness', Sober notes, is supervenient on a wide disjunction of different packages of environment-organism relations, it is multiply realized by them. Nevertheless, he insists:

> ['fitness's] irreducible utility arises from its contribution to the construction of theories that allow us to subsume [that is, explain] a huge variety of single events within a common framework. This explanatory virtue of the fitness concept is a consequence of that property's *supervenience*.[9]

And for good measure, Sober extends the point to psychology: "Science does need to attribute abstract mental states to organisms if it wishes to carve out generalizations at that level of abstraction we can in principle predict and explain your behavior from the vantage point of your current mental state and also from the vantage point of your current physical state" (*ibid.*, p. 128).

In a series of papers stretching from 1984 to 1999, Philip Kitcher[10] has argued for the same conclusions, employing in particular the example of pair-separation in meiosis as a complete, adequate, and irreducible explanation of genetic recombination in spite of the fact that meiosis is a multiply realized functional kind. In its most recent version, the claim takes the following form: consider explanandum

> (G) Genes on different chromosomes, or sufficiently far apart on the same chromosome, assort independently.

According to Kitcher,[11] there is an explanation for G that employs the functional taxonomy of classical genetics which proceeds as follows:

> (PS) Consider the following kind of process, a *PS*-process (for *pairing* and *separation*). There are some basic entities that come in pairs. For each pair, there is a correspondence relation between the parts of one member of the pair and the parts of the other member. At the first stage of the process, the entities are placed in an *arena*. While they are in the arena, they can exchange segments, so that the parts of one member of a pair are replaced by the corresponding parts of the other members, and conversely. After exactly one round of exchanges, one and only one member of each pair is drawn from the arena and placed in the *winners box*.

---

[9] *The Nature of Selection*, p. 126.

[10] "1953 and All That: A Tale of Two Sciences," *Philosophical Review*, xciii (1984): 335-73.

[11] "The Hegemony of Molecular Biology," *Biology and Philosophy*, xiv (1999): 199-200.

> In any PS-process, the chances that small segments which belong to members of different pairs or which are sufficiently far apart on members of the same pair will be found in the winners box are independent of one another (G) holds because the distribution of chromosomes to games at meiosis is a PS-process
>
> This I submit is a full explanation of (G), and explanation that prescinds entirely from the stuff that genes are made of (*ibid*, pp 199-200)

As Kitcher had previously argued "Antireductionists are not only able to contend that there are autonomous levels of biological explanation They can also resist the weaker reductionist view that explanation always flows from the molecular level up "[12]

There are two conclusions to draw here one bears on Shapiro's proposal for dealing with his "dilemma", the other is wider Shapiro's proposal will not do, for it in effect condemns special sciences to an eliminativist fate Shorn of an explanatory role, once their usefulness in taxonomizing explananda has been exhausted, they have no further function The real moral of the fact that, in Shapiro's words, functional kinds "never enter into (nonanalytic) laws of the form 'All *P*s are also ___'," is that they are nomological danglers (654) It is for this reason that Shapiro is ultimately impaled on horn 2, eliminativism

The wider lesson recalls the alternative suggestion about how to slip between the horns of Shapiro's dilemma At the beginning of this section, I noted that we could avoid reductionism and eliminativism if explanation in the special sciences proceeded without the explicit participation of laws The trouble with this approach is that it requires a whole new account of how the special sciences explain In the end, it appears, the obligation to provide such an account will be unavoidable For it is the same obligation incurred by the invocation of 'temporary', historically limited nonnomological reductions as explanatory which was discussed in section I above

Shapiro is right that functional kinds never enter into (nonanalytic) laws of the form 'All *Ps* are also ___' But in the absence of laws of the form 'All *P*s are also ___', exponents of the autonomy of the special sciences will need a new account of explanation, one which provides for their *unaided* explanatory power in the absence of laws expressed in their proprietary taxonomies

ALEX ROSENBERG

Duke University

---

[12] "1953 and All That," p 371

BOOK REVIEWS

Liberalism, Constitutionalism, and Democracy. RUSSELL HARDIN. New York: Oxford University Press, 1999. xviii + 379 p. Cloth $29.95.

Rich in historical detail and philosophical nuance, Russell Hardin's *Liberalism, Constitutionalism, and Democracy*'s central thesis is that constitutional politics constitute an "unequal coordination game." It is a "coordination game" in the standard sense: every player prefers any outcome within the set of "coordinated solutions" to any feasible outcome outside that set. It is "unequal" in that players' preferences differ within that set.

This is familiar to game theorists as the "Battle of the Sexes." I prefer going to the opera rather than the mud wrestling, my partner prefers the opposite; but above all, we both prefer doing something together rather than going our separate ways.

Hardin thinks constitutional politics are like that, and offers the United States as a case in point. The Federalists preferred a stronger central state, Anti-Federalists a weaker one. But once there was sufficient support for the Federalist plan, then it was clear that the choice was between that model or none at all, it was in even the Anti-Federalists' interests to acquiesce, rather than form a small and exposed state adjacent to a large and powerful union capable of overwhelming it, both militarily and economically. Edmund Randolph said as much, in explaining why he voted at the Virginia ratifying convention for the Constitution he had refused to sign at Philadelphia (111).

Unequal coordination models accommodate the critical legal theorists' insight that institutional structures and rights regimes are (almost) never neutral, while at the same time accommodating the conventionalists' observation that (almost) everyone is (almost) always better off abiding by whatever rules everyone else is abiding by. Unequal coordination also explains the bandwagon dynamics of constitutional politics. Everyone starts out fervently pushing his own preferred plan, but once some particular solution has emerged as the clear favorite, everyone tends suddenly to fall into line up behind that—which makes perfect sense, assuming each group weakly prefers one coordinated solution over others but strongly prefers any coordinated solution over none.

Other consequences of this model are less pleasing, not least perhaps to Hardin himself. First, it relegates democracy to the status

0022-362X/01/9807/374–78

of a sideshow. "Democracy works only when there are no deeply divisive issues" (1) of, for example, an ethnic sort (276ff). Democracy always deals only with "chaff at the margins" (30), after all the core issues have been constitutionally settled. Presented provocatively as a point about democracy and constitutionalism-cum-coordination, this is actually a more mundane truth concerning the powerlessness of day-to-day politics (democratic or otherwise) where higher-order politics remain unsettled Just as there is no reason to buy title to property if it is cheaper to buy the judges and police charged with enforcing those titles, so too there is no reason to legislate under a constitution when it is easier to amend the constitution directly or to substitute a new one. Constitutions have to be genuinely entrenched, genuinely hard to change, before ordinary politics (democratic or otherwise) can come into play

Second, coordination sometimes settles on solutions that are not only "unequal" (differentially advantaging some over others) but also "partial" (advantaging others not at all, or even leaving them worse off than in an uncoordinated state) (3-4) So long as the group is sufficiently small or marginalized not to be able seriously to disrupt the existing constitutional settlement, it can be left out of the constitutionalist calculations of mutual advantage

Third, "in advanced liberal democracies in our time, the groups whose mutual advantage *must* be served are the large middle classes and the wealthy—especially, among the latter, the entrepreneurially wealthy . Liberal governments do not commonly fall because of the opposition of the poor" (5, 7) Constitutional politics in eighteenth-century America are thus interpreted as a conspiracy of northern commercial and southern plantation interests in free trade triumphing over small-holder agrarian Anti-Federalist interests Political and economic liberalization in eastern Europe in our own day is viewed through the same "commercial order" spectacles

Although Hardin treats this as a real-world observation which needs to be imported into his model, the power of commercial interests can actually be linked directly into Hardin's unequal coordination logic. Simply recall the dynamics of coordination and the strategic advantages of organization Commercial interests are relatively well organized, purely for commercial purposes in the first instance, but with considerable political consequences in the second Their superior organization makes them quicker and more effective at building coalitions around their preferred coordination solution And once some particular solution begins to emerge as the most salient, recall, everyone in a coordination game (even an unequal one) tends to fall into line behind it.

(This analysis also incidentally offers the poor a counterstrategy If only they could organize—first themselves, and then others—around their own preferred coordination solution, theirs would be the solution toward which the bandwagon dynamics of coordination would tip But just try to imagine the poor out-organizing the rich!)

Analytically, Hardin is at pains to insist that constitution making is a coordination game, not a contract game The difference is one of strategic structure, prisoner's dilemma games (in which you would prefer to defect if others cooperate) versus coordination games (in which you would not) The prisoner's dilemma structure governs exchange and contract relations, because there is a first-choice preference 'I want everything I have plus everything you have—and you do, too' Hardin says that that is an inappropriate model for constitution making

> What makes the problem of coordination rise above the conflict in my wanting everything and your wanting everything is the tremendous prospect for production and mutual gain from allowing each of us to keep some of what we have and produce If we successfully coordinate on a governmental structure to establish order and regulate exchange, our outcome is not then worse for me than the state in which I get everything we already have and you get nothing It is radically better than that outcome It is an outcome that only a productive society could provide for any of us (97)

Note well, however to turn the structure of preferences underlying constitution making into a coordination game rather than a prisoner's dilemma, those productivity gains must be very great—so great as to make my payoff under the postconstitution regime larger than the payoff of 'I get everything we already have and you get nothing' under the preconstitution regime And of course to make it a coordination game from everyone's perspective, that must be true not just of me and you but of everyone else as well Extrapolating from the two-person (me and you) case to the *N*-person case of a real society, the productivity gains must indeed be so large that each and every one of us gets more postgovernment than all *N* of us got pregovernment Then and only then would each of us prefer, as our first-choice option, cooperating in a world with a constitution to taking everything everyone else had (if only we could!) in a world without a constitution

But where in the world has instituting government ever caused the per capita national gross domestic product even so much as to double overnight (the two-person version of the story), much less to be multiplied by a factor equivalent to the population itself (the *N*-

person version)? Substantial growth might come eventually, over the course of many generations (though even then, never by that extraordinary multiple) But even that will get no grip on players of Hardin's games, unless they have in effect a massively negative discount rate (taking more—very substantially more—interest in distant generations' welfare than their own)

Of course, it is pragmatically crazy to fixate on the gains of the 'I defect, you cooperate' cell Realistically, no one can seriously expect to seize everything anyone has for herself Realistically, the only outcomes that have much hope of being remotely stable in prisoner's dilemma are the symmetrical ones both defect, or both cooperate And if the choice is among just that pair of options, then ex hypothesi we are both better off both cooperating

But that is merely to say unilateral welching on an on-going basis is simply not in the feasible set (not "workable," in the sociological sense to which Hardin often appeals) What makes it infeasible in repeated prisoner's dilemma games is that others' cooperation is conditional on ours, and if we welch then (purely as a matter of prudential self-protection) they will defect as well What makes it infeasible where prisoner's dilemma games are resolved by the sword of sovereign is the credible threat of external enforcement of the contract (although again, the sword of sovereign derives its own strength from the backing of the general public)

Once external enforcements are in place, of course, the strategic structure of the game has changed, and the game is no longer one of prisoner's dilemma at all That, as I understand it, is Hardin's point Constitution making in the first instance is a coordination game In the second instance, once a government has been established with power to impose penalties and sanctions, constitution keeping is not a prisoner's dilemma, either Instead, it is a game in which cooperating (abiding by the law) is a strictly dominant strategy There is no moment of prisoner's dilemma, and hence no contract story to be told, anywhere along the line

There is indeed an important difference between negotiating (or renegotiating) a contract or constitution in the first instance and abiding by one once one is in place The former is indeed a coordination game But contrary to the sharp contrast Hardin wants to draw, that is as true of contract making as of constitution making The economist's "contract curve" is a line representing a veritable infinity of points each of which is preferable from both contractor's perspective to nonagreement Contract making is an unequal coordination game, because some of those potential terms of contract are better than others for each player But negotiating terms of contract is definitely a coordination game for both players, because nonagreement means no gain for anyone Ditto for negotiating constitutions, just as Hardin says

Abiding by the terms of an agreement, once one is in place, typically has the strategic structure of a prisoner's dilemma, however That is as true for constitution keeping as for contract keeping Just as each party to a multilateral contract would be better off if everyone else abided by it while she alone was free to deviate, so too would each party to a constitutional settlement be better off if everyone else abided by the constitution while doing as they pleased

Surely, for example, it would have been in the Federalists' interests to retain office after losing the 1800 election, had they been able to count on everyone *else* following the Constitution faithfully regardless (cf 136-37) John Adams vacated office after his arch-enemy Thomas Jefferson defeated him, simply because that was the only way for the Federalists to continue enjoying the benefits of the Constitution at all, including secure possession of the Presidency if they ever rewon it. (Even had they forseen they never would, constitutions are package deals, and the only way Federalists could continue enjoying the advantages of free trade among the states that the Constitution ensured was to comply with the Constitution in its various other particulars)

Where everyone clearly anticipates retaliation and the consequent unravelling of the constitutional settlement, and everyone clearly dreads that outcome, it will standardly be the case that no major player gives any serious consideration to behaving in an unconstitutional manner But that is not to say that threats of retaliation were absent or irrelevant, or that the constitutional settlement prevailed because no one was tempted to defect (as in Hardin's self-enforcing coordination game) Nor is it to say that major players are compelled by the sword of sovereign (which, after all, would have remained in President Adams's hand had he refused to quit office) Rather, it is to say that constitution keeping is ordinarily just a prisoner's dilemma game played among a sufficiently small set of constitutionally significant actors for cooperation to be credibly conditional on the performances of all concerned

ROBERT E GOODIN

Australian National University

NEW BOOKS

ABBEY, RUTH *Nietzsche's Middle Period* New York Oxford, 2000 xvii, 208 p Cloth $45 00

AINSLIE, GEORGE *Breakdown of the Will* New York Cambridge, 2001 xi, 258 p Cloth $54 95, paper $19 95

ALLISON, DAVID B *Reading the New Nietzsche* The Birth of Tragedy, The Gay Science, Thus Spoke Zarathustra, *and* On the Genealogy of Morals Lanham, MD Rowman and Littlefield, 2001 xvi, 316 p Cloth $75 00, paper $24 95

ALLISON, HENRY E *Kant's Theory of Taste A Reading of the* Critique of Aesthetic Judgment New York Cambridge, 2001 xvi, 424 p Cloth $69 95, paper $24 95

ALSTON, WILLIAM P *A Sensible Metaphysical Realism* Milwaukee. Marquette, 2001 65 p Cloth $15 00

ANTON, COREY *Selfhood and Authenticity* Albany SUNY, 2001 x, 181 p Cloth $49 50, paper $16 95

AVRAMIDES, ANITA *Other Minds* New York Routledge, 2001 xv, 333 p Cloth $75 00, paper $24 95

AUDI, ROBERT *The Architecture of Reason The Structure and Substance of Rationality* New York Oxford, 2001 xvi, 286 p Cloth $35 00

AZZOUNI, JODY *Knowledge and Reference in Empirical Science* New York Routledge, 2000 xi, 259 p Cloth $85 00

BAILLIE, JAMES *Routledge Philosophy Guidebook to Hume on Morality* New York Routledge, 2000 ix, 226 p Cloth $60 00

BARKER, ANDREW *Scientific Method in Ptolemy's* Harmonics New York Cambridge, 2000 viii, 281 p Cloth $69 95

BARRY, BRIAN *Culture and Equality An Egalitarian Critique of Multiculturalism* Cambridge Harvard, 2001 xi, 399 p Cloth $35 00

BARTLETT, ROBERT C *The Idea of Enlightenment A Post-Mortem Study* Toronto University Press, 2001 xi, 224 p Cloth $45 00

BEAUCHAMP, TOM L, and JAMES F CHILDRESS *Principles of Biomedical Ethics,* fifth edition New York Oxford, 2001 xi, 454 p Cloth $49 95

BENCIVENGA, ERMANNO *Exercises in Constructive Imagination* Boston Kluwer, 2001 vii, 211 p Cloth $89 00

BENNINGTON, GEOFFREY *Interrupting Derrida* New York Routledge, 2000 xiv, 235 p Cloth $75 00, paper $24 99

BIRD, ALEXANDER *Thomas Kuhn* Princeton University Press, 2000 xii, 308 p Cloth $39 50, paper $16 95

BIRO, MATTHEW *Anselm Keifer and the Philosophy of Martin Heidegger* New York Cambridge, 1998 xiv, 319 p Paper $27 95

BOSTOCK, DAVID *Aristotle's Ethic* New York Oxford, 2000 ix, 255 p Cloth $24 95

BRANN, EVA *The Ways of Naysaying No, Not, Nothing, and Nonbeing* Lanham, MD Rowman and Littlefield, 2001 xviii, 249 p Cloth $35 00

BRENNAN, TERESA *Exhausting Modernity Grounds for a New Economy* New York Routledge, 2000 viii, 216 p Cloth $65 00, paper $19 95

BROWNE, GREGORY M *Necessary Factual Truth* Lanham, MD UP of America, 2001 xxii, 388 p Cloth $46 00

CALHOUN, CHESHIRE *Feminism, the Family, and the Politics of the Closet Lesbian and Gay Displacement* New York Oxford, 2000 vii, 172 p Cloth $29 00

CARROLL, NOEL *Beyond Aesthetics Philosophical Essays* New York Cambridge, 2001 xiv, 450 p Paper $27 95

CARRUTHERS, PETER *Phenomenal Consciousness A Naturalistic Theory* New York Cambridge, 2000 xx, 347 p Cloth $59 95

CATALANO, JOSEPH S *Thinking Matter* New York Routledge, 2000 vii, 224 p Paper $20 00

CHARLES, DAVID *Aristotle on Meaning and Essence* New York Oxford, 2001 xv, 410 p Cloth $55 00

CHEETHAM, MARK A *Kant, Art, and Art History Moments of Discipline* New York Cambridge, 2001 x, 222 p Cloth $54 95

CLARK, ANDY *Mindware An Introduction to the Philosophy of Cognitive Science* New York Oxford, 2000 xiii, 210 p Paper $18 95

CLARKE, W NORRIS, S J *The One and the Many A Contemporary Thomistic Metaphysics* Notre Dame University Press, 2001 viii, 324 p Cloth $45 00, paper $24 00

CLAY, DISKIN *Platonic Questions Dialogues with the Silent Philosopher* University Park Penn State UP, 2000 xxiii, 309 p Cloth $40 00, paper $17 95

COHEN, G A *Karl Marx's Theory of History A Defence, Expanded Edition* Princeton University Press, 2000 xxviii, 442 p Paper $24 95

COLYVAN, MARK *The Indispensability of Mathematics* New York Oxford, 2001 viii, 172 p Cloth $35 00

CONWAY, DAVID *The Rediscovery of Wisdom From Here to Antiquity in Quest of Sophia* New York St Martin's, 2000 x, 196 p Cloth $69 95

COOLE, DIANA *Negativity and Politics Dionysus and Dialectics from Kant to Poststructuralism* New York Routledge, 2000. x, 272 p Cloth $85 00, paper $25 99

CORMIER, HARVEY *The Truth Is What Works William James, Pragmatism, and the Seed of Death* Lanham, MD Rowman and Littlefield, 2001 xv, 189 p Cloth $63 00, paper $19 95

CORRINGTON, ROBERT S *A Semiotic Theory of Theology and Philosophy* New York Cambridge, 2000 xi, 268 p Cloth $59 95

CRAIG, WILLIAM LANE *The Tenseless Theory of Time A Critical Examination* Boston Kluwer, 2000 x, 256 p Cloth $103 00, *The Tensed Theory of Time A Critical Examination* Boston Kluwer, 2000 x, 287 p Cloth $123 00, *Time and the Metaphysics of Relativity* Boston Kluwer, 2001 xi, 279 p Cloth $92 00

CREAVEN, SEAN *Marxism and Realism A Materialistic Application of Realism in the Social Sciences* New York Routledge, 2000 x, 325 p Cloth $110 00

CROUCH, MARGARET A *Thinking about Sexual Harasment A Guide for the Perplexed* New York Oxford, 2001 315 p Cloth $60 00, paper $19 95

CROWELL, STEVEN GALT *Husserl, Heidegger, and the Space of Meaning Paths toward Transcendental Phenomenology* Evanston, IL Northwestern, 2001 xviii, 323 p Cloth $79 95, paper $27 95

CURREN, RANDALL R *Aristotle on the Necessity of Public Education* Lanham, MD Rowman and Littlefield, 2000 xv, 278 p Cloth $69 00, paper $24 95

DAINTON, BARRY *Stream of Consciousness Unity and Continuity in Conscious Experience* New York Routledge, 2000 xvi, 254 p Cloth $75 00

DANCY, JONATHAN *Practical Reality* New York Oxford, 2000 xii, 187 p Cloth $35 00

DAVIES, PAUL SHELDON *Norms of Nature Naturalism and the Nature of Functions* Cambridge MIT, 2001 xiv, 234 p Cloth $35 00

DELEDALLE, GÉRARD *Charles S Peirce's Philosophy of Signs Essays in Comparative Semiotics* Bloomington Indiana UP, 2001 xii, 199 p Cloth $39 95

DENIS, LARA *Moral Self-Regard Duties to Oneself in Kant's Moral Theory* New York Routledge, 2001 xvi, 245 p Cloth $75 00

DES CHENE, DENNIS *Life's Form Late Aristotelian Conceptions of the Soul* Ithaca Cornell, 2000 viii, 220 p Cloth $45 00, *Spirits and Clocks Machine and Organism in Descartes* Ithaca Cornell, 2001 xiii, 181 p Cloth $39 95

DESMOND, WILLIAM *Ethics and Between* Albany SUNY, 2001 xiii, 530 p Paper $29 95

EARMAN, JOHN *Hume's Abject Failure The Argument against Miracles* New York Oxford, 2000 xi, 217 p Cloth $39 95, paper $21 95

EATON, MARCIA MUELDER *Merit, Aesthetic and Ethical* New York Oxford, 2001 vii, 252 p Cloth $49 95

EBBS, GARY *Rule-Following and Realism* Cambridge Harvard, 2001 xvi, 368 p Paper $24 95

EDGLEY, ALISON *The Social and Political Thought of Noam Chomsky* New York Routledge, 2000 viii, 205 p Cloth $85 00

ELDRIDGE, RICHARD *The Persistence of Romanticism Essays in Philosophy and Literature* New York Cambridge, 2001 xii, 251 p Cloth $54 95, paper $19 95

ELLSBERG, DANIEL *Risk, Ambiguity, and Decision* New York Routledge, 2001 liii, 281 p Cloth $65 00

ESKIN, MICHAEL *Ethics and Dialogue* New York Oxford, 2001 xvi, 294 p Cloth $85 00

FARRAR, ROXANNE CLAIRE *Sartrean Dialectics A Method for Critical Discourse on Aesthetic Experience* Atlanta Rodopi, 2000 xiii, 101 p Paper $21 00

FERRARIN, ALFREDO *Hegel and Aristotle* New York Cambridge, 2001 xxii, 442 p Cloth $64 95

FERRÉ, FREDERICK *Living and Value Toward a Constructive Postmodern Ethics* Albany SUNY, 2001 xix, 364 p Cloth $78 50, paper $26 95

FLEW, ANTHONY *Merely Mortal? Can You Survive Your Own Death?* Amherst, NY Prometheus, 2000 xviii, 200 p Cloth $26 00

FORTH, CHRISTOPHER E *Zarathustra in Paris The Nietzsche Vogue in France, 1891-1918* DeKalb Northern Illinois UP, 2001 ix, 238 p Cloth $42 00

FOTION, NICK *John Searle* Princeton University Press, 2000 v, 266 p. Paper $16.95.

FREADMAN, RICHARD *Threads of Life Autobiography and the Will* Chicago: University Press, 2001 xiii, 394 p Cloth $55 00, paper $19 00

FREELAND, CYNTHIA *But Is It Art? An Introduction to Art Theory* New York Oxford, 2001 xxi, 231 p Cloth $16.95

FRENCH, PETER A : *The Virtues of Vengeance* Lawrence Kansas UP, 2001 xii, 248 p Cloth $29 95

FRIEDLANDER, ELI *Signs of Sense A Reading of Wittgenstein's* Tractatus Cambridge Harvard, 2001 xx, 227 p Cloth $42 50

FRIEDMAN, MICHAEL *A Parting of the Ways Carnap, Cassirer, and Heidegger* Chicago Open Court, 2000. xv, 175 p Paper $24.95

FYNSK, CHRISTOPHER: *Infant Figures: The Death of the 'Infans' and Other Scenes of Origin* Stanford University Press, 2000. ix, 199 p Cloth $49 50, paper $18 95

GAITA, RAIMOND. *A Common Humanity. Thinking About Love and Truth* New York: Routledge, 2000 xxxiii, 293 p Cloth $27 50

GANSON, DORIT A *The Explanationist Defense of Scientific Reason* New York: Routledge, 2001 xv, 206 p Cloth $60.00

GARBER, DANIEL *Descartes Embodied: Reading Cartesian Philosophy through Cartesian Science* New York. Cambridge, 2001 xii, 337 p Cloth $64.95, paper $22 95

GARBER, MARJORIE *Academic Instincts.* Princeton University Press, 2001 xi, 187 p Cloth $19 95

GAUKROGER, STEPHEN *Francis Bacon and the Transformation of Early-Modern Philosophy* New York. Cambridge, 2001 xii, 249 p Cloth $59 95, paper $21 95

GENDLER, TAMAR SZABÓ: *Thought Experiment: On the Powers and Limits of Imaginary Cases* New York Routledge, 2000 xvii, 258 p Cloth $75 00

GILLIES, DONALD: *Philosophical Theories of Probability* New York Routledge, 2000 xiv, 223 p Cloth $75 00, paper $24 99.

GILMORE, JONATHAN *The Life of a Style Beginnings and Endings in the Narrative History of Art.* Ithaca. Cornell, 2000 xiv, 157 p Cloth $35 00

GOLDIE, PETER *The Emotions A Philosophical Exploration* New York. Oxford, 2000. vi, 265 p Cloth $35 00

GORDON, PAUL. *Tragedy after Nietzsche Rapturous Superabundance.* Urbana Illinois UP, 2001 162 p Cloth $29.95.

GREENBERG, ROBERT *Kant's Theory of A Priori Knowledge.* University Park Penn State UP, 2001 ix, 278 p. Cloth $45 00

GUTTING, GARY. *French Philosophy in the Twentieth Century* New York: Cambridge, 2001 xiv, 419 p. Cloth $64 95, paper $24 95

HAHN, LEWIS E *A Contextualistic Worldview.* Carbondale: Southern Illinois UP, 2001. xviii, 187 p. Cloth $39 95

HAHN, ROBERT *Anaximander and the Architects The Contributions of Egyptian and Greek Architectural Technologies to the Origins of Greek Philosophy* Albany SUNY, 2001 xxiii, 326 p Paper $27 95

HANEY, DAVID P *The Challenge of Coleridge Ethics and Interpretation in Romanticism and Modern Philosophy* University Park Penn State UP, 2001 xviii, 309 p Cloth $55 00

HANNA, ROBERT *Kant and the Foundations of Analytic Philosophy* New York Oxford, 2001 xv, 312 p Cloth $65 00

HANSSEN, BEATRICE *Critique of Violence Between Poststructuralism and Critical Theory* New York Routledge, 2000 vi, 314 p Cloth $80 00, paper $25 95

HARMAN, GILBERT *Explaining Value and Other Essays in Moral Philosophy* New York Oxford, 2000 xiv, 238 p Cloth $55 00, paper $19 95

HARRE, ROM *One Thousand Years of Philosophy From Ramanuja to Wittgenstein* Malden, MA. Blackwell, 2000 xix, 362 p Cloth $59 95

HARRIES, KARSTEN *Infinity and Perspective* Cambridge MIT, 2001 xii, 380 p Cloth $37 95

HEATH, JOSEPH *Communicative Action and Rational Choice* Cambridge, MIT, 2001 xii, 363 p Cloth $39 95

HEISIG, JAMES W *Philosophers of Nothingness An Essay on the Kyoto School.* Honolulu Hawai'i UP, 2001 xi, 380 p Paper $21 95

HICKMAN, LARRY A *Philosophical Tools for Technological Culture Putting Pragmatism to Work.* Bloomington Indiana UP, 2001 xi, 215 p Cloth $39 95, paper $17 95

HILL, RUTH *Sceptres and Sciences in the Spains Four Humanists and the New Philosophy (ca 1680-1740)* Liverpool University Press, 2000 vii, 296 p Cloth $56 95

HONDERICH, TED *Philosopher A Kind of Life* New York Routledge, 2001 x, 441 p Cloth $30 00

HOOKER, BRAD *Ideal Code, Real World A Rule-Consequentialist Theory of Morality* New York Oxford, 2000 xiii, 213 p Cloth $35 00

HORNSBY, JENNIFER *Simple Mindedness In Defense of Naive Naturalism in the Philosophy of Mind* Cambridge Harvard, 2001 ix, 265 p Paper $19 95

HORTY, JOHN F *Agency and Deontic Logic* New York Oxford, 2001 xi, 192 p Cloth $29 95

HUME, DAVID *An Enquiry concerning Human Understanding* TOM L BEAUCHAMP, ed New York Oxford, 2001 civ, 344 p Cloth $65 00

HUNTER, IAN *Rival Enlightenments Civil and Metaphysical Philosophy in Early Modern Germany* New York Cambridge, 2001 xix, 398 p Cloth $69 95

HURKA, THOMAS *Virtue, Vice, and Value* New York Oxford, 2001 ix, 272 p Cloth $55 00

HYDE, MICHAEL J *The Call of Conscience Heidegger and Levinas, Rhetoric and the Euthanasia Debate* Columbia South Carolina UP, 2001 xvii, 300 p Cloth $39 95

IANNONE, A PABLO *Dictionary of World Philosophy* New York Routledge, 2001 xviii, 554 p Cloth $140 00

JANIK, ALLAN *Wittgenstein's Vienna Revisited* New Brunswick, NJ Transaction, 2001 xv, 287 p Cloth $49 95

JARDINE, NICHOLAS *The Scenes of Inquiry* New York Oxford, 2000 xix, 317 p Paper $24 95

JOHNSON, ROLF M *Three Faces of Love* Dekalb Northern Illinois UP, 2001 x, 208 p Cloth $36 00

JOLLIMORE, TROY A *Friendship and Agent-Relative Morality* New York Routledge, 2001 xiii, 176 p Cloth $65 00

JOPLING, DAVID A *Self-Knowledge and the Self* New York Routledge, 2000 ix, 193 p Cloth $58 00, paper $18 95

JURIST, ELLIOT *Beyond Hegel and Nietzsche Philosophy, Culture, and Agency* Cambridge MIT, 2001 xii, 355 p Cloth $37 95

KADVANY, JOHN *Imre Lakatos and the Guises of Reason* Durham, NC Duke, 2001 xx, 378 p Cloth $69 95, paper $23 95

KEEFE, ROSANNA *Theories of Vagueness* New York Cambridge, 2000 xii, 233 p Cloth $54 95

KEIJZER, FRED *Representation and Behavior* Cambridge MIT, 2001 x, 276 p Cloth $40 00

KELLEY, DAVID *The Contested Legacy of Ayn Rand Truth and Toleration in Objectivism* New Brunswick, NJ Transaction, 2000 131 p Paper $18 95

KELLY, SEAN D *The Relevance of Phenomenology to the Philosophy of Language and Mind* New York Routledge, 2001 xii, 176 p Cloth $65 00

KING, JERRY C *Complex Demonstratives A Quantificational Account* Cambridge MIT, 2001 xiii, 206 p Cloth $45 00, paper $18 00

KOLAK, DANIEL *On Hintikka* Belmont, CA Wadsworth, 2001 ii, 88 p Paper $14 95

KOONS, ROBERT C *Realism Regained An Exact Theory of Causation, Teleology, and the Mind* New York Oxford, 2001 xvi, 349 p Cloth $45 00

KRAUSZ, MICHAEL *Limits of Rightness* Lanham, MD Rowman and Littlefield, 2000 ix, 169 p Cloth $60 00, paper $22 95

KUEHN, MANFRED *Kant A Biography* New York Cambridge, 2001 xxii, 544 p Cloth $34 95

KUKLA, ANDRE *Social Constructivism and the Philosophy of Science* New York Routledge, 2000 x, 170 p Paper $25 99

LAMB, DAVID *The Search for Extraterrestrial Intelligence A Philosophical Inquiry* New York Routledge, 2001 x, 210 p Cloth $65 00, paper $19 95

LENNOX, JAMES G *Aristotle's Philosophy of Biology Studies in the Origin of Life Science* New York Cambridge, 2001 xxiii 321 p Cloth $64 95, paper $22 95

LEVINE, JOSEPH *Purple Haze The Puzzle of Consciousness* New York Oxford, 2001 204 p Cloth $29 95

LOPSTON, PETER *Reality Fundamental Topics in Metaphysics* Toronto University Press, 2001 xii, 305 p Cloth $65 00

LUGG, ANDREW *Wittgenstein's Investigations 1-133 A Guide and Interpretation* New York Routledge, 2000 ix, 214 p Cloth $75 00

MADISON, G B , PAUL FAIRFIELD, and INGRID HARRIS *Is There a Canadian Philosophy? Reflections on the Canadian Identity* Ottawa University Press, 2000 vii, 218 p Paper $28 00

MAGNANI, LORENZO *Abduction, Reason, and Science Processes of Discovery and Explanation* New York Kluwer, 2001 xvii, 205 p Cloth $85 00

MANDEVILLE, BERNARD *Free Thoughts on Religion, The Church, and National Happiness* IRWIN PRIMER, ed New Brunswick, NJ Transaction, 2001 xxxv, 234 p Cloth $49 95

MARKS, SUSAN *The Riddle of All Constitutions International Law, Democracy, and the Critique of Ideology* New York Oxford, 2000 163 p Cloth $49 95

MARTIN, BILL *The Radical Project Sartrean Investigations* Lanham, MD Rowman and Littlefield, 2000 x, 127 p Cloth $59 95, paper $19 95

MAY, TODD *Our Practices, Our Selves, or, What It Means to Be Human* University Park Penn State UP, 2001 x, 206 p Cloth $35 00, paper $16 95

MAYBERRY, J P *The Foundations of Mathematics in the Theory of Sets* New York Cambridge, 2001 xx, 424 p Cloth $80 00

MCCARNEY, JOSEPH *Hegel on History* New York Routledge, 2000 xi, 234 p Cloth $50 00, paper $12 99

MCCUMBER, JOHN *Time in the Ditch American Philosophy and the McCarthy Era* Evanston, IL Northwestern, 2001 xxiii, 213 p Cloth $29 95

MCEVOY, JAMES *Robert Grosseteste* New York Oxford, 2000 xx, 219 p Cloth $35 00, paper $18 95

MCINERNY, RALPH *Characters in Search of Their Author* Notre Dame University Press, 2001 xii, 138 p Cloth $25 00

MCKENZIE, MICHAEL C *Paul Ramsey's Ethics The Power of 'Agape' in a Postmodern World* Westport, CT Praeger, 2001 xxii, 173 p Cloth $60 00

MCGINN, COLIN *Logical Properties Identity, Existence, Predication, Necessity, Truth* New York Oxford, 2001 vi, 114 p Cloth $24 95

MELE, ALFRED, R *Self-Deception Unmasked* Princeton University Press, 2001 xi, 148 p Cloth $49 50, paper $15 95

MENSCH, JAMES RICHARD *Postfoundational Phenomenology Husserlian Reflections on Presence and Embodiment* University Park Penn State UP, 2001 275 p Cloth $45 00

MERCER, DAVID E *Kierkegaard's Living-Room The Relation between Faith and History in* Philosophical Fragments Montreal McGill-Queen's, 2001 ix, 207 p Cloth $60 00

MEWS, CONSTANT J *The Lost Love Letters of Heloise and Abelard* With translations by NEVILLE CHIAVAROLI and CONSTANT J MEWS New York Palgrave, 2001 xvii, 378 p Paper $18 95

MIDGLEY, MARY *Science and Poetry* New York Routledge, 2001 x, 230 p Cloth $30 00

MOHANTY, J N *The Self and Its Other Philosophical Essays* New York Oxford, 2000 xv, 146 p Cloth $24 95

MONK, RAY *Bertrand Russell The Ghost of Madness, 1921-1970* New York Free Press, 2000 xv, 574 p Cloth $40 00

NATHAN, N M L *The Price of Doubt* New York Routledge, 2001 viii, 177 p Cloth $75 00

NELSON, HILDE LINDEMANN *Damaged Identities, Narrative Repair* Ithaca Cornell, 2001 xv, 204 p Cloth $42 50, paper $17 95

NOONAN, HAROLD W *Frege A Critical Introduction* Malden, MA Blackwell, 2001 x, 244 p Cloth $59 95

NORRIS, CHRISTOPHER *Deconstruction and the 'Unfinished Project of Modernity'* New York Routledge, 2000 240 p Cloth $85 00, paper $24 95

O'DEA, JANE *Virtue or Virtuosity? Explorations in the Ethics of Musical Performance* Westport, CT Greenwood, 2000 xii, 122 p Cloth $55 00

ODIN, STEVE *Artistic Detachment in Japan and the West Psychic Distance in Comparative Aesthetics* Honolulu Hawai'i UP, 2001 vii, 294 p Paper $29 95

O'NEILL, ONORA *Bounds of Justice.* New York Cambridge, 2000 ix, 219 p Cloth $54 95, paper $19 95

OSBORNE, PETER *Philosophy in Cultural Theory* New York Routledge, 2000 xiii, 146 p Cloth $65 00, paper $20 95

OWENS, DAVID *Reason without Freedom The Problem of Epistemic Normativity* New York Routledge, 2000 viii, 199 p Cloth $90 00, paper $29 99

PATTISON, GEORGE *The Later Heidegger* New York Routledge, 2000 xiii, 230 p Cloth $60 00, paper $14 99

PATTISON, STEPHEN *Shame Theory, Therapy, Theology* New York Cambridge, 2000 x, 343 p Cloth $64.95, paper $22 95

PERRY, JOHN *Knowledge, Possibility, and Consciousness* Cambridge MIT, 2001 xvi, 221 p Cloth $29 95

PETERS, CARMICHAEL C *A Gadamerian Reading of Karl Rahner's Theology of Grace and Freedom* New York Catholic Scholars, 2000 xi, 373 p Cloth $64 00, paper $44 50

PHILLIPS, D C *The Expanded Social Scientist's Bestiary A Guide to Fabled Threats to, and Defenses of, Naturalistic Social Science* Lanham, MD Rowman and Littlefield, 2000 xi, 277 p Cloth $72 00, paper $31 95.

PIERCE, CHRISTINE *Immovable Laws, Irresistible Rights Natural Law, Moral Rights, and Feminist Ethics* Lawrence Kansas UP, 2000 x, 171 p Cloth $29 95

POLKA, BRAYTON *Depth Psychology, Interpretation, and the Bible An Ontological Essay on Freud* Montreal McGill-Queen's, 2001 xviii, 397 p Cloth $70 00

PORTER, JAMES I *Nietzsche and the Philology of the Future* Stanford University Press, 2000 xiii, 449 p Cloth $60 00, paper $19 95

PRADHAN, R C *Recent Developments in Analytic Philosophy* New Delhi Indian Council of Philosophical Research, 2001 xx, 516 p

PREUS, J SAMUEL *Spinoza and the Irrelevance of Biblical Authority* New York Cambridge, 2001 xvi, 228 p Cloth $54 95

PRIEST, STEPHEN *The Subject in Question· Sartre's Critique of Husserl in* The Transcendence of the Ego New York Routledge, 2000 ix, 181 p Cloth $85 00

PROOPS, IAN *Logic and Language in Wittgenstein's* Tractatus New York. Garland, 2000 xxii, 121 p Cloth $55 00

RAABE, PETER B *Philosophical Counseling Theory and Practice* Westport, CT Praeger, 2001 xxii, 303 p Cloth $64 95

RAMPLEY, MATTHEW *Nietzsche, Aesthetics, and Modernity* New York Cambridge, 2000 xi, 286 p Cloth $59 95

RAPPE, SARA *Reading Neoplatonism Non-discursive Thinking in the Texts of Plotinus, Proclus, and Damascius* New York Cambridge, 2000 xxii, 266 p Cloth $59 95

REGAN, TOM *Defending Animal Rights* Urbana Illinois UP, 2001 xii, 179 p Cloth $24 95

RESCHER, NICHOLAS. *Process Philosophy A Survey of Basic Issues* Pittsburgh University Press, 2000 vii, 144 p. Cloth $27 50

ROSENBERG, ALEX *Philosophy of Science A Contemporary Introduction* New York Routledge, 2000 x, 191 p Cloth $65 00, paper $19 99

ROSS, STEPHEN DAVID *The Gift of Property Having the Good* Albany SUNY, 2001 xvii, 347 p Cloth $78 50, paper $26 95.

SACKS, MARK *Objectivity and Insight* New York Oxford, 2000 346 p Cloth $55 00

SANTER, ERIC L *On the Psychotheology of Everyday Life Reflections on Freud and Rosenzweig* Chicago· University Press, 2001 viii, 156 p Cloth $40 00, paper $16 00

SEDGWICK, PETER *Descartes to Derrida An Introduction to European Philosophy* Malden, MA. Blackwell, 2001 xi, 310 p Paper $29 95

SHADE, PATRICK *Habits of Hope A Pragmatic Theory* Nashville Vanderbilt, 2001 xv, 238 p Cloth $34 95

SIM, STUART *Post-Marxism An Intellectual History* New York Routledge, 2000. 198 p Cloth $85 00

SIMMONS, A JOHN *Justification and Legitimacy Essays on Rights and Obligations* New York Cambridge, 2001 xiv, 276 p Cloth $54 95, paper $19 95

SLOTE, MICHAEL *Morals from Motives* New York Cambridge, 2001 xv, 216 p Cloth $39 95

SON, CHANG-HEE *Haan of Minjung Theology and Han of Han Philosophy* Lanham, MD UP of America, 2000 xvii, 157 p Cloth $49 00, paper $29 50

SPEIGHT, ALLEN *Hegel, Literature, and the Problem of Agency* New York Cambridge, 2001 xii, 154 p Cloth $54 95, paper $18 95

STELLARDI, GIUSEPPE *Heidegger and Derrida on Philosophy and Metaphor. Imperfect Thought* Amherst, NY Humanity, 2000 286 p Paper $27 00

STERELNY, KIM *The Evolution of Agency and Other Essays* New York Cambridge, 2001 xvi, 310 p Cloth $54 95, paper $19 95

STERN, JOSEF *Metaphor in Context*, Cambridge MIT, 2000 xvii, 385 p Cloth $39 95

STRATTON-LAKE, PHILIP *Kant, Duty, and Moral Worth* New York Routledge, 2000 xi, 153 p Cloth $75 00

STURGEON, SCOTT *Matters of Mind Consciousness, Reason, and Nature* New York Routledge, 2000 x, 166 p Cloth $85 00, paper $27 99

SZABÓ, ZOLTÁN GENDLER *Problems of Compositionality* New York Garland, 2000 xi, 165 p Cloth $55 00

TAMEN, MIGUEL *The Matter of the Facts On Invention and Interpretation* Stanford University Press, 2000 168 p Cloth $45 00

TAN, KOK-CHOR *Toleration, Diversity, and Global Justice* University Park Penn State UP, 2000 xii, 233 p Cloth $28 50

TARASTI, EERO *Existential Semiotics* Bloomington Indiana UP, 2000 218 p Cloth $39 95

TAUBER, ALFRED I *Henry David Thoreau and the Moral Agency of Knowing* Berkeley California UP, 2001 xi, 317 p Cloth $40 00

THAGARD, PAUL *Coherence in Thought and Action* Cambridge MIT, 2000 xiv, 312 p Cloth $32 95

THOMSON, JUDITH JARVIS *Goodness and Advice* AMY GUTMANN, ed Princeton University Press, 2001 xvi, 188 p Cloth $19 95

TILES, J E *Moral Measures An Introduction to Ethics West and East* New York Routledge, 2000 xii, 334 p Cloth $75 00, paper $20 99

TOMASI, JOHN *Liberalism beyond Justice Citizens, Society, and the Boundaries of Political Theory* Princeton University Press, 2001 xviii, 163 p Cloth $55 00, paper $15 95

TOULMIN, STEPHEN *Return to Reason* Cambridge Harvard, 2001 x, 243 p Cloth $24 95

TOWNSEND, DABNEY *Hume's Aesthetic Theory Taste and Sentiment* New York Routledge, 2001 x, 258 p Cloth $85 00

VALADEZ, JORGE M *Deliberative Democracy, Political Legitimacy, and Self-Determination in Multicultural Societies* Boulder, CO Westview, 2001 xiv, 386 p Paper $40 00

VELLEMAN, J DAVID *The Possibility of Practical Reason* New York Oxford, 2000 viii, 302 p Paper $19 95

VOGLER, CANDACE A *John Stuart Mill's Deliberative Landscape An Essay in Moral Psychology* New York Routledge, 2001 xvi, 136 p Cloth $65 00

WALLACH, JOHN R *The Platonic Political Art A Study of Critical Reason and Democracy* University Park Penn State UP, 2001 xi, 468 p Cloth $65 00, paper $25 00

WARBURTON, NIGEL *Freedom An Introduction with Readings* New York Routledge, 2001 viii, 252 p Cloth $65 00, paper $20 99, *Thinking from A to Z,* second edition New York Routledge, 2000 xi, 150 p Paper $50 00

WEHRLE, WALTER E *The Myth of Aristotle's Development and the Betrayal of Metaphysics* Lanham, MD Rowman and Littlefield, 2000 xiii, 279 p Cloth $67 50, paper $24 95

WEISS, ROSLYN *Virtue in the Cave Moral Inquiry in Plato's* Meno New York Oxford, 2001 x, 229 p Cloth $39 95

WELTON, DONN *The Other Husserl The Horizons of Transcendental Phenomenology* Bloomington Indiana UP, 2001 xvi, 496 p Cloth $54 95

WILLIAMS, LINDA L *Nietzsche's Mirror The World as Will to Power* Lanham, MD Rowman and Littlefield, 2001 xiv, 149 p Cloth $65 00, paper $22 95

WITTENBERG, DAVID *Philosophy, Revision, Critique Rereading Practices in Heidegger, Nietzsche, and Emerson* Stanford University Press, 2001 xiii, 266 p Cloth $45 00

WOOD, DAVID *The Deconstruction of Time* Evanston, IL Northwestern, 2001 xxxiv, 431 p Paper $24 95

YOUNG, JULIAN *Heidegger's Philosophy of Art* New York Cambridge, 2001 xiii, 179 p Cloth $49 95

THE JOURNAL OF PHILOSOPHY
VOLUME XCVIII, NO. 8, AUGUST 2001

NEUTRALITY AND RESPONSIBILITY

The principle that the state should be neutral among different conceptions of the good life is a much discussed topic in contemporary political philosophy. Defenders of this principle[1] have argued that only a neutral state can show equal respect to all persons in a pluralistic democratic society. Critics of this principle[2] have argued that a neutral state is neither necessary nor sufficient for a well-functioning democratic society, and that a neutral state would unjustifiably fail to promote valuable options and discourage disvaluable ones.

Yet neither defenders nor critics of the principle of state neutrality have sufficiently discussed the extent to which the state is responsible for what it does and for what it allows to happen.[3] To be sure, an understanding of the state's responsibilities underlies different characterizations of the neutral state, but this understanding is seldom openly identified and discussed. I shall address this issue head on. I present the case for thinking that the state is negatively responsible for what it allows to happen as well as positively responsible for what it does. And I draw out the implications that this expanded view of state responsibility has for the claim that the state should be neutral among different conceptions of the good life. My discussion exposes

---

[1] See Bruce Ackerman, *Social Justice in the Liberal State* (New Haven: Yale, 1980), Charles Larmore, *Patterns of Moral Complexity* (New York: Cambridge, 1987), and Thomas Nagel, "Moral Conflict and Political Legitimacy," *Philosophy and Public Affairs*, XVI (Summer 1987): 215-40.

[2] See Joseph Raz, *The Morality of Freedom* (New York: Oxford, 1986), George Sher, *Beyond Neutrality* (New York: Cambridge, 1997); and my *Liberalism, Perfectionism, and Restraint* (New York: Cambridge, 1998).

[3] I write here as if the state were a moral agent. This is done for convenience of exposition. I do not mean to deny (or affirm) that all of the responsibilities of the state are reducible to the responsibilities of individuals.

0022-362X/01/9808/389-410

an important tension between the principle of state neutrality, on its most widely accepted understanding, and strongly egalitarian conceptions of social justice Overcoming this tension, I argue, significantly raises the costs of accepting this principle

I TWO CONCEPTIONS OF STATE NEUTRALITY

I begin with a general statement of the principle of state neutrality Then I describe two competing interpretations of this principle, both of which have been widely discussed in the philosophical literature As we shall see, these two interpretations rest on (or are supported by) different understandings of the state's responsibilities

Understood very generally, the principle of state neutrality holds that the state[4] should not favor (or disfavor) any permissible non-neutral conception of the good or give greater (or lesser) assistance to those who pursue it This statement of the principle contains numerous ambiguities, but only two of them need concern us here The first ambiguity concerns the qualifier 'permissible' According to the principle of state neutrality, the state must be neutral, not among all actual, but among all permissible conceptions of the good While the principle of state neutrality is a principle of political morality, it is not the only such principle Some actual conceptions of the good may be impermissible because they conflict with these other principles For example, a conception of the good that includes the domination of some over others might be impermissible because it conflicts with a principle that holds that all persons have equal moral worth

The principle of state neutrality applies only to the class of permissible conceptions of the good, however this class might be delimited I shall assume that there exists a wide range of permissible conceptions of the good, even though this range is smaller than the range of actual conceptions of the good [5]

A second, and more important, ambiguity in the above statement concerns how we should understand the claim that the state should not *favor* any non-neutral conception of the good Depending on how this claim is construed, we get different characterizations of the principle of state neutrality This brings me to the two common interpretations alluded to above

---

[4] To avoid cumbersome qualifications, I shall refer to "the state" to mean either the constitutional framework sustained by the modern state *or* that framework plus the laws and policies passed within it Most of the time I shall have the latter, more inclusive, idea in mind But my arguments go through even if the principle of neutrality is restricted to the constitutional framework

[5] Henceforth, I shall speak simply of conceptions of the good, but this should be understood as a shorthand for permissible conceptions of the good

(A) The state should not aim to do anything to promote any non-neutral conception of the good, or give greater assistance to those who pursue it, unless a plausible neutral justification can be given for the state's action

(B) The state should not do anything or refrain from doing anything that has the effect—whether intended or not—of promoting any non-neutral conception of the good, or of providing greater assistance to those who pursue it

Those who favor interpretation (A) hold that the principle of state neutrality requires the state only not to aim to advantage any non-neutral conception of the good [6] Those who favor interpretation (B) hold that this is not sufficient They think the state should do all it can to ensure that the effects of its actions and inactions do not advantage or disadvantage any non-neutral conception of the good

Sophisticated proponents of the principle of state neutrality have almost universally rejected interpretation (B) Nearly everyone of them[7] has embraced interpretation (A), which I shall henceforth refer to as *neutrality of justification,* over interpretation (B) which I shall henceforth refer to as *neutrality of effect* In the main, this latter view has been rejected because it is thought to be too demanding It appears to be either impossible to achieve or, if it is achievable, to require a much more interventionist state than many would be comfortable with Despite this, I intend to show that some of the assumptions regarding the state's responsibilities which underlie the consensus view that neutrality of justification is a better interpretation of the principle of state neutrality than neutrality of effect have not been sufficiently thought through and that, when they are thought through, the consensus view will look less plausible

What, then, are these assumptions? To identify them I need to introduce two distinct, but overlapping, distinctions about responsibility These distinctions are common fare in discussions of personal responsibility,[8] but they are much less commonly applied to state responsibility The two distinctions, respectively, draw on the difference between doing (or bringing about) and allowing, and the difference between intending and merely foreseeing Applied to the state, we can ask Is the state responsible only for what it does or also for what it allows[9] to happen? And we can ask Is the state responsible only for the intended consequences of its interventions or also for the unintended consequences as well? More to the issue at hand, we can ask When the state intervenes or fails to intervene in ways that affect the ability of persons to pursue different conceptions of the good, is it responsible only for what it does and for the intended consequences of its interventions or is it responsible for more?

---

[6] There can be cases where state officials aim to favor a non-neutral conception of the good, but where their action is nonetheless neutrally justifiable These cases ask us to think about what is fundamental to interpretation (A)—the neutral justifiability of state action, the intention of state officials, or both This has led some writers to distinguish neutrality of justification from neutrality of aim I shall not be concerned with these cases here, however I shall assume artificially that, if state action is neutrally justifiable, then those who undertake this action do not aim to favor or disfavor a non-neutral conception of the good

[7] There are a few exceptions See Robert Goodin and Andrew Reeve, "Do Neutral Institutions Add Up to a Neutral State?" in Goodin and Reeve, eds , *Liberal Neutrality* (New York Routledge, 1989), pp 193-210, and David Miller, *Market, State, and Community* (New York Oxford, 1990), pp 72-97

Proponents of neutrality of justification have not squarely confronted these questions, but it seems clear that they must limit the state's responsibility in various ways to prevent their view from collapsing into neutrality of effect The proponent of neutrality of effect will claim that what matters are the results of state action or inaction When the state fails to intervene when it could have done so, then it is responsible for the advantages or disadvantages to different conceptions of the good which its failure to intervene allows Likewise, when the state's intervention produces unintended consequences that advantage or disadvantage different conceptions of the good, then it is responsible for these consequences But, surely, the proponent of neutrality of justification will object that this makes the state responsible for too much The neutral state, he will claim, is a state that undertakes action only for neutral reasons It is not a state that is morally responsible for every advantage or disadvantage that accrues to every proponent of every conception of the good which exists within its jurisdiction Such a state would be unduly intrusive It would always be interfering with one side or the other to make sure all sides were treated fairly

At first blush, then, it appears that the differences between neutrality of justification and neutrality of effect track differences in underlying assumptions about the state's responsibilities, and that these underlying assumptions are based on the familiar distinctions in moral philosophy between doing and allowing and intending and merely foreseeing As we shall see later, this is not entirely correct Proponents of neutrality of justification may be able to limit the state's responsibilities in other ways by appealing to different distinctions, distinctions which are related to the ones we have discussed but are not identical with them But for now it should be plain that the two conceptions of state neutrality I have identified rest on different accounts of the state's responsibilities An investigation into which conception is superior, therefore, must address this issue

---

[8] For thorough, but critical, discussion of these distinctions, see Shelly Kagan, *The Limits of Morality* (New York Oxford, 1989), pp 83-182

[9] The state "allows" an event to happen if and only if it does not prevent the event from happening and it was in the state's power to prevent it from happening

II CONSEQUENTIAL RESPONSIBILITY

The state is not responsible for everything that happens within its jurisdiction Many things that happen within its jurisdiction are beyond its control Other things that happen are within its control, but the state should not be faulted for allowing them to happen How then are we to decide what the state is and is not responsible for?

Rather than attempt to answer this large question, I want to address a smaller, but related, question What kinds of *outcomes* is the state responsible for? Let us say that an outcome is a consequence that results from an intervention that the state undertakes or a consequence that is the result of the state's failure to undertake an intervention that it could have undertaken So described, an outcome can be the result of the state's decision to intervene or its decision not to intervene, but it must be something that it is plausible to think that it could have made a decision about[10] The state is not responsible for the occurrence of a hurricane, but it may be responsible for the consequences that result from its failure to issue warnings to residents in an area hit by a hurricane

The kinds of outcomes I am concerned with here are outcomes that have a significant impact on the ability of persons to pursue their conception of the good An outcome is neutral if it neither advantages nor disadvantages any non-neutral conception of the good, and it is non-neutral if it does advantage or disadvantage a non-neutral conception of the good More precisely, then, the question I want to address is What kinds of non-neutral outcomes is the state responsible for?

With this in mind, it will be helpful to distinguish between six types of non-neutral outcomes

NNO(1) A non-neutral consequence that is the intended result of state intervention

NNO(2) A non-neutral consequence that is the intended result of the state's failure to intervene when it could have intervened in a way that would have prevented the consequence

---

[10] Strictly speaking, outcomes do not result from the state's decision to intervene or not to intervene, but from the state's intervention or nonintervention This is true since a decision to intervene may not be actually carried out When I speak of the outcome that results from the state's decision to act, I should be understood to mean the consequence that results when the decision is actually carried out

NNO(3) A non-neutral consequence that is the unintended, but foreseeable, result of state intervention

NNO(4) A non-neutral consequence that is the unintended, and unforeseeable, result of state intervention

NNO(5) A non-neutral consequence that is the unintended, but foreseeable, result of the state's failure to intervene when it could have intervened in a way that would have prevented the consequence

NNO(6) A non-neutral consequence that is the unintended, and unforeseeable, result of the state's failure to intervene when it could have intervened in way that would have prevented the consequence

The descriptions of the first three of these outcome types are reasonably clear, but NNO(4), NNO(5), and NNO(6) require further comment When the state fails to intervene, this may be because it has decided not to intervene or it may be because it never even considered intervening, which may be true of NNO(5) and NNO(6) outcomes The state, however, can be responsible for an outcome even if it never considered the option of intervening in a way that would have prevented it This will be true in cases where it is both plausible to think that it could have intervened and that it should have done so In the hurricane example above, the state is responsible for the outcome that results from its failure to issue warnings even if it never considered doing so

Furthermore, the claim that the state is responsible for NNO(4) and NNO(6) outcomes is stronger, and less plausible, than the claim that it is responsible for the other types of outcomes distinguished here How, it may be asked, can the state be held responsible for outcomes it neither intended nor could have foreseen? The answer is that while it cannot be faulted for the decisions which it made which resulted in these outcomes, it nonetheless can be responsible for rectifying or compensating for them If the state could not have foreseen the occurrence of the hurricane, then it could not have foreseen the effects of its failure to issue a warning Still, it may be responsible for compensating those who were victims of the hurricane

In general, I shall assume that the state is responsible for a non-neutral outcome that results from its interventions (or failure to intervene when it could have done so) if there is at least a *pro tanto* reason for it to compensate those who have been disadvantaged by the outcome This means that the state can be responsible for a non-neutral outcome even if it does not have a conclusive moral obligation to compensate those who have been disadvantaged This will occur in contexts where the costs to others of compensating those who have been disadvantaged override the *pro tanto* reason to compensate

I said above that proponents of neutrality of justification must find a way to limit the state's responsibility They do this by focusing on what the state *aims* to bring about It should now be clearer what this means Since the state can aim to bring about an outcome either by deciding to intervene or by deciding not to intervene, it seems clear that proponents of neutrality of justification will need to hold that the state is responsible for NNO(1) and NNO(2) types of outcomes And since NNO(3), NNO(4), NNO(5), and NNO(6) types of outcomes do not result from the state's intentional decisions, it seems clear that they will need to hold that the state is not responsible for them

III SOCIAL JUSTICE AND THE COMPARATIVE-POSSIBILITIES APPROACH

Thus far I have been discussing the state's responsibility for non-neutral outcomes But the question of the state's responsibility for outcomes also arises in the context of social justice Indeed, the most prominent proponents of neutrality of justification combine their defense of neutrality with a particular understanding of social justice Since this understanding of social justice is independently plausible and since it explicitly raises the question of what the state is and is not responsible for, we need to consider it here

The understanding I have in mind puts the focus on the basic structure[11] of a self-contained, large modern society It holds that the institutions that comprise this structure are just if and only if there is not some alternative feasible set of institutions for this society which is more fair or more acceptable to the people who live in it Thus, on this understanding, the institutions that comprise the basic structure can be judged to be unjust because of outcomes they in fact bring about or prevent as well as outcomes that those who set them up and participate in them aim to bring about or prevent For example, if a group of people were in a very bad economic position, not because the state intentionally and actively intervened in order to put them there, but because the state simply failed to enact measures that it could have enacted, then the state would be responsible for their condition Whether the state's failure to enact these measures would be unjust would depend on whether the measures, if enacted, would excessively disadvantage other groups.

---

[11] This term is taken from John Rawls, of course, and I am using it in the same way that he uses it

This understanding of social justice is an instance of what I shall call the *comparative-possibilities approach*. According to this approach, political and economic institutions as well as laws and policies passed by the state, must be justified by comparison with all other possible feasible institutional/policy options.[12] When applied to issues of social justice, this approach has been used to support a fairly strong form of egalitarianism.[13] Crucially, this strong form of egalitarianism rests on the idea that the state is responsible both for what it does and for what it allows to happen. As Thomas Nagel explains:

> [T]he acceptance of a serious egalitarian ideal would have to appeal to a notion of negative responsibility, on the part of society, for failing to arrange things differently in ways that it could. ... Noninterference requires justification as much as interference does. *Every* arrangement has to be justified by comparison with every other real possibility.[14]

In addition, although Nagel does not mention this, strong egalitarianism is also supported by the idea that the state is responsible both for what it intentionally brings about and for the unintended consequences of its interventions. This is true because, if a group of persons were disadvantaged by the unintended consequences of state intervention, then, on this view, the state would be responsible for their condition, so long as there were some alternative institutional/policy arrangement that did not have these consequences and did not unduly disadvantage any other group.

Thus, when issues of social justice are being considered—and in particular when the distribution of economic resources is being considered—strong egalitarianism downplays the moral significance of the distinctions between what the state does and what it allows to happen and what the state intends to bring about and what it actually brings about. This does not imply that these distinctions lose all moral significance. Strong egalitarians may still claim that it is more offensive if the state brings about intentionally a state of affairs in which a group of persons suffer unduly than if it allows this to happen unintentionally But, according to the comparative-possibilities approach, this kind of distinction is not morally fundamental What matters on this approach is how different groups of people fare under different possible institutional/policy arrangements, and to make these judgments it is not necessary to refer to the distinctions we have been discussing[15]

---

[12] See Rawls, *A Theory of Justice* (Cambridge: Harvard, 1971), and Thomas Nagel, *Equality and Partiality* (New York: Oxford, 1991).

[13] By itself, of course, the comparative-possibilities approach does not lead to strong egalitarianism, but when it is combined with a special concern for the worst-off group, then it does so. Henceforth, when I speak of strong egalitarianism I shall have in mind a view that combines the comparative-possibilities approach with a special concern for the worst-off group. In doing this, I do not mean to deny that there may be other egalitarian views that do not rest on this approach.

[14] *Equality and Partiality*, p. 100.

I shall say, then, that according to strong egalitarianism, the distinctions between *what the state does* and *what it allows to happen* and the distinction between *what the state aims to bring about* and *what it actually brings about* do not have *direct* moral significance[16] It may turn out that these distinctions have indirect moral significance for strong egalitarians Suppose, for example, that it is true that, if state officials were to accept these distinctions as morally fundamental, then they would be more likely to make decisions that made the worst-off group better off than if they did not accept these distinctions as morally fundamental Granting this supposition, the strong egalitarian could argue that the best institutional/policy arrangement would be one in which these distinctions were recognized as morally fundamental This would establish that the distinctions have *indirect* moral significance

But, at least at first blush, it is hard to see why these distinctions would have indirect moral significance for the strong egalitarian As some have pointed out, these distinctions actually benefit the better-off at the expense of the worst-off by limiting the moral responsibility of the former to help the latter[17] Still, deeper investigation may reveal that everyone is made better off by a state whose members accept and act on these distinctions For now, I want to leave this possibility open Strong egalitarianism, as I shall understand it, is committed only to the view that the distinctions we have been discussing, when applied to issues of social justice, do not have direct moral significance

---

[15] In downplaying the significance of these distinctions, the comparative-possibilities approach to social justice has a consequentialist flavor As I have indicated, however, it is an approach to social justice that is best developed in the work of Rawls This requires comment Rawls's conception of justice falls short of a consequentialist conception by (a) giving greater weight to the needs and interests of the worst-off compared to the better-off, and (b) limiting its concern to resource, as opposed to environmental, advantages and disadvantages As I have made plain, embracing (a) makes the comparative-possibilities approach strongly egalitarian Below I explain and critically discuss (b) For a helpful discussion of the "consequentialism" of Rawls's conception of justice, see Thomas Pogge, *Realizing Rawls* (New York Columbia, 1989), pp 36-47

[16] This is subject to the qualification noted in the previous paragraph Many strong egalitarians would likely hold that intentional state intervention that places a person in a bad economic position is more offensive than state neglect that unintentionally leaves him in the same bad position If so, then the distinctions I have been focusing on would have some direct moral significance for them But, since this qualification does not affect my general argument, I shall put it aside

[17] See Gilbert Harman's contribution to Harman and Judith Jarvis Thomson, *Moral Relativism and Moral Objectivity* (Cambridge Blackwell, 1996), pp 24-25

The question I now wish to raise is Does strong egalitarianism, as I have characterized it, sit well with neutrality of justification? Are these two views compatible, or must we choose between them? This is an important question, for if it turns out that these two views are incompatible, then a strong argument, one that draws on the persuasiveness of strong egalitarianism, could be mounted against neutrality of justification

IV THE ASYMMETRY

As we have seen, the principle of state neutrality holds that the state must be neutral among conceptions of the good life This principle does not hold that the state must be neutral among different conceptions of social justice Given that the principle of neutrality is itself a moral principle, one that is a constituent part of a conception of justice, a state that recognizes and acts on this principle would not be neutral toward those conceptions of social justice which rejected it

This, of course, should not embarrass the neutralist He believes we can and should draw a distinction between social justice and conceptions of the good life The principle of neutrality applies to the latter, not the former Working with this distinction, the neutralist can claim that what the state is and is not responsible for varies according to the kind of issues involved More specifically, he can say that when issues of social justice are in play, one set of ideas applies, and when issues concerning the good life are in play, a different set applies This thought, if sense can be made of it, would allow the strong egalitarian to embrace neutrality of justification He could claim that when issues of social justice are on the table, the distinctions we have been discussing have no direct moral significance, but when issues concerning the good life are on the table, these distinctions come into their own and have direct moral significance

But is this thought coherent? Can we distinguish between issues of social justice and issues of the good life in the way it suggests? People can be advantaged and disadvantaged by the state in myriad ways If we think that some of these ways are relevant to social justice, while others are not, then how do we determine this?

To think through these matters, it will help to have some concrete examples before us Consider these four

(A) The state represses a particular religion It does so with the aim of making it more likely that this religion will die off Its action is successful As a result of its repressive measures, membership in the religion declines significantly

(B) The state undertakes an aggressive civic-education program in its public schools This program is justified on the grounds that it will cultivate political skills and civic commitments that will enable all students to participate effectively in the democratic political process when they become adults The program is successful, but it has the unintended, albeit foreseeable, side effect of making it significantly more difficult for a particular traditional religious community to survive and prosper The members of this community wish to exempt their children from the civic-education program, but this would conflict with the program's goal of training all young people to be future citizens

(C) Members of a democratically organized workers' cooperative ask the state for public money to help finance the investment needed if their firm is to remain in business They point out that their firm is not being run poorly, but that it is structurally disadvantaged by the competitive market economy [18] Without public money, their firm will not survive The state decides that considerations of economic efficiency take precedence over the need to preserve particular firms It therefore does not provide the public money As a result, the workers' cooperative firm goes under and the workers are forced to look for work in noncooperative firms

(D) An economic program, consisting of public investment and redistributive policies, is proposed If adopted, the program would raise the income of those who currently have less than one-half of the median income (and there is no other feasible economic program that could achieve this goal equally well) State officials consider the program, but reject it on the grounds that it would hinder economic growth The program is not undertaken by the state

The proponent of neutrality of justification, as I have described him, will hold that the state is not responsible for the disadvantages to non-neutral conceptions of the good which result from its action, or failure to act, in cases (B), (C), and (D) The reason for this is that in (B) the disadvantages are unintended side effects of the state's action, whereas in (C) and (D) the disadvantages are the unintended consequence of the state's failure to act [19] If the state is responsible only for the consequences it aims to brings about, then it is not responsible for the disadvantages suffered by the groups in these three cases It is clear, however, that the strong egalitarian will want to say that the state is responsible for the disadvantages that its failure to act creates in (D) This would remain true even if the state's failure to act in (D) were neutrally justified

---

[18] This claim is standardly advanced by defenders of market socialism See Miller, *Market, State, and Community*, pp 92-97, and Jon Elster, "From Here to There Or If Co-operative Ownership Is So Desirable, Why Are There So Few Co-operatives?" *Social Philosophy and Policy*, vi (Spring 1989) 93-111 For the purposes of this example, I assume this claim is correct

I have suggested that this need not embarrass the proponent of neutrality of justification who is also a strong egalitarian He need not choose between his commitment to state neutrality and his commitment to strong egalitarianism, for he can insist that the principle of state neutrality does not apply to conceptions of social justice It applies only to conceptions of the good, and the relevant disadvantages in (D) engage social justice But, we must now ask Why do the relevant disadvantages in (B) and (C) not also engage social justice? If we say they do, then the distinction between neutrality of justification and neutrality of effect will be in trouble If just any disadvantage to any conception of the good that the state engenders by its action or inaction can engage social justice, and if the principle of neutrality does not apply to issues of social justice, then neutrality of justification will collapse into neutrality of effect If, on the other hand, only some kinds of disadvantages to conceptions of the good engage issues of social justice, then we need to be given some explanation for why this is so

Let us consider more carefully, then, the differences among these four cases In (A), the state intervenes in order to set back the interests of those who pursue a particular conception of the good [20] It does so by restricting the religious liberty of these people In (D), the state fails to act in such a way to ensure that a disadvantaged economic group has a higher share of economic resources Here, the ability of these people to pursue their conception of the good is set back by the state's failure to act, at least to the extent that economic

[19] More precisely, and using the terminology introduced above, case (B) is an example of an NNO(3) outcome, while cases (C) and (D) are examples of an NNO(5) outcome

[20] It might be said that the state acts wrongly in this case not because it acts non-neutrally, but because it violates individual rights This may be correct, but even if it is, the neutralist will have an additional reason for judging the state's action to be wrong

resources enable them to pursue their conception of the good In both of these cases, then, some people are disadvantaged by the state's action (or its failure to act) and the disadvantages imposed take the form of either an infringement of liberty or a reduced share of economic resources I shall refer to these kinds of disadvantages as *resource disadvantages*

In contrast, in (B) and (C), the disadvantages suffered by the relevant groups do not take the form of either an infringement of liberty or a reduced share of economic resources Rather, to the extent that the relevant groups in these cases have a complaint, it is that they are living in a social/institutional environment that is not conducive to the successful pursuit of their conception of the good They can argue that the unintended side effects of the state's action in (B) or the failure of the state to intervene in (C) has contributed to a social/institutional environment that imposes disadvantages on them I shall refer to these kind of disadvantages as *environmental disadvantages*

The distinction between these two kinds of disadvantages brings into view an important line of argument open to the proponent of neutrality of justification Social justice, it can be said, concerns only the proper assignment and distribution of resources With respect to resource advantages and disadvantages, the state is responsible both for what it does and for what it allows to happen It also is responsible for the unintended side effects of its interventions But, it can be further said, the state is not responsible for all the disadvantages that persons suffer as a result of the social/institutional environment that the state sustains Here the principle of state neutrality comes into play to limit its responsibility The principle holds that the state is responsible only for the (non-neutral) environmental disadvantages that it intentionally brings about

This line of argument reconciles neutrality of justification with the comparative-possibilities approach to social justice It does so by pressing hard on the distinction between resource and environmental disadvantages It maintains that there is an asymmetry between the state's responsibility for the former and the state's responsibility for the latter Henceforth, I shall refer to this simply as the *asymmetry*

Can the asymmetry be grounded? Is there an independent justification for holding the state more responsible for resource disadvantages than for environmental disadvantages, or has the asymmetry been introduced solely to reconcile neutrality of justification with strong egalitarianism?

V STATE NEUTRALITY AND PERSONAL RESPONSIBILITY

To address these questions, I now consider what I take to be the strongest argument in support of the asymmetry[21] This argument draws on an attractive picture of people's responsibility for their own ends It can be set out as follows

(i) The state's responsibility for non-neutral outcomes is limited by what it is reasonable or appropriate to hold individual subjects responsible for
(ii) It is reasonable or appropriate to hold individual subjects responsible for the environmental disadvantages they suffer in a neutrally justified state that has a socially just distribution of resources
(iii) Therefore, when individual subjects suffer environmental disadvantages in such a state, the state is not responsible for these disadvantages and hence the individual subjects have no claim for compensation against the state

I shall refer to this as the *argument from personal responsibility* The first premise in the argument looks to be correct If an individual is (entirely) responsible for a given outcome, then he cannot hold the

---

[21] Naturally, there are other arguments that some will think provide support for the asymmetry Two of them warrant mention here First, it is often claimed that neutrality of effect is practically impossible and that therefore it must be rejected This claim can be developed in a way that supports the asymmetry The argument would be that while it is possible for the state to be neutral with respect to the decisions which it makes which affect the share of resources that its citizens have, it is not possible for it to be neutral with respect to the decisions which it makes which sustain a social/institutional environment When it is making the latter type of decisions, any decision it makes will have non-neutral consequences Accordingly, so it may be concluded, the state cannot be held responsible for the environmental disadvantages its citizens suffer—unless perhaps it has aimed to bring them about in order to disadvantage particular citizens

But this conclusion does not follow For the argument does not really establish that neutrality of effect must be rejected Like neutrality of justification, neutrality of effect can and should be understood to be a target to be aimed at So long as it is possible to get closer to the target (I shall have more to say on this below), neutrality of effect will be a workable principle, one capable of guiding state policy

The second argument holds that so long as the state does not intentionally favor some conceptions of the good over others, the social/institutional environment that it sustains will be one that best promotes the good of all its members According to this argument, the environment that results from the state's benign neglect of different conceptions of the good will be like a free-competitive market in which good conceptions drive out bad This, in turn, suggests that when we are considering environmental disadvantages, as opposed to resource disadvantages, the distinction between what the state does and what it allows to happen and the distinction between the consequences the state aims to bring about and those it merely foresees have indirect moral significance They have indirect moral significance because all people, irrespective of the content of their conception of the good, will be better off if they are recognized In this way, the argument appears to provide support for the asymmetry state responsible for that outcome This explains why those who have cultivated expensive tastes cannot object to equal shares of resources on the grounds that equality would make it harder for them to satisfy their tastes compared to those with less expensive tastes[22] If we are responsible for the tastes we cultivate, then we cannot hold the state responsible for the difficulty we encounter in trying to satisfy them

Of course, in many contexts this judgment will not be so easy Quite often, the responsibility for an outcome is shared For example, suppose one cultivates an interest in opera Suppose further that the state imposes a tax on opera houses because it wishes to discourage the enjoyment of opera Here one may be able to hold the state partially responsible for the difficulty one has in satisfying one's tastes One may have a claim to compensation because the state has acted so as to bring about a non-neutral outcome Still, this kind of case raises a difficult question If we know the state is taxing opera houses, then why are we not responsible for this extra expense when we cultivate the taste for it? Why should the tax not be considered one further factor contributing to an environment in which the enjoyment of opera is expensive? Presumably, the proponent of neutrality of justification will argue that, if the tax is neutrally justified, then it should be considered part of the environment against which individuals make decisions about which tastes to cultivate But

---

The appearance, however, is misleading There is ample reason to reject the central claim of this argument In a social/institutional environment that results from benign neglect, good ways of life might fare poorly This could be true for a variety of reasons Most straightforwardly, people might have inadequate information about or insufficient appreciation of various valuable options Simply ensuring that they are free to make choices does not guarantee that they will make good choices Less obviously, but equally true, some valuable conceptions of the good might fare poorly under conditions of benign neglect Given prevailing preference structures, and given the efficiency demands of the economy, some conceptions of the good might not survive or flourish in the competitive environment of benign neglect This is not to say that in this environment bad conceptions of the good would drive out good ones It is to say only that some valuable conceptions of the good would not fare well This is sufficient to undercut the argument For if it is true that some valuable conceptions of the good would not fare well under benign neglect, then it will not be true that this environment is one that all persons, irrespective of the content of their conception of the good, have good reason to support Some persons will be able to object that an environment of benign neglect disadvantages them in their efforts to pursue their conception of the good Whether this objection will be able to justify a claim to compensation will depend on the costs that such compensation would impose on others But this argument does not give us any reason to think that this kind of compensation could never be justified

[22] Ronald Dworkin, "What Is Equality? Part 1 Equality of Welfare," *Philosophy and Public Affairs*, x (Summer 1981) 185-246

if it is not neutrally justified, if it is imposed for the purpose of setting back the interests of those who like opera, then the state is responsible for the disadvantage this imposes on those who cultivate this taste.

This brings me to the second premise in the argument from personal responsibility. For this argument to ground the asymmetry, it must substantiate the distinction between resource and environmental disadvantages. Strong egalitarians can argue that, while the state is responsible for ensuring that each citizen has an equal or fair share of resources, each individual is responsible for the use he makes of these resources. On this view, people have no moral claim against the state to live in an environment that is hospitable to their preferred conception of the good. Rather, they have a responsibility to adjust their aims to the social/institutional environment sustained by the state that results from the free choices of their fellow citizens.

If correct, this line of thought would show that in a socially just and neutrally justified state, individuals are responsible for the environmental disadvantages they confront. Granting the first premise of the argument from personal responsibility, this would license the conclusion that people can have no moral claim for compensation against the state for these disadvantages. This, in turn, would ground the asymmetry, thus providing a non-ad-hoc reconciliation of neutrality of justification with strong egalitarianism.

But here it is necessary to ask: Why should individuals be held responsible for the environmental disadvantages they confront whenever they have a fair share of resources? It is sometimes suggested that our conceptions of the good, like our preferences and tastes, lie within our voluntary control. We have the power to adjust them to our circumstances and so we are responsible for doing so. But this response is unpersuasive for two reasons. First, as many have pointed out,[23] it is not true that our tastes, preferences, and conceptions of the good are always under our voluntary control. Many of the aims and aspirations that shape our lives were never chosen by us and some of them we may not be able to abandon or alter. Second, this response does not identify a reason for treating resource disadvantages differently from environmental disadvantages. In a free society the resource shares that we have are determined in part by our voluntary choices and decisions. If I squander my money at the race tracks, then I shall have fewer resources than you because of the voluntary choices I have made. So if we appeal to voluntary choice to account for the disadvantages that we are and are not responsible for, then this will not support the second premise of the argument we are considering. Rather, it will support the idea that the important division between disadvantages is that between voluntary and involuntary disadvantage,[24] not that between resource and environmental disadvantage. And this, of course, does not help at all in grounding the asymmetry.

---

[23] See Scanlon, "The Significance of Choice," in Sterling M. McMurrin, ed., *The Tanner Lectures on Human Values, Volume VIII* (Salt Lake City: Utah UP, 1988), pp. 149-216; G. A. Cohen, "On the Currency of Egalitarian Justice," *Ethics*, XCIX (July 1989): 906-44; and Richard Arneson, "Primary Goods Reconsidered," *Noûs*, XXIV (June 1990): 429-54.

If we cannot appeal to the voluntary character of the preferences, aims, and aspirations that comprise our conceptions of the good, then how can we account for our responsibility for them? Perhaps the answer lies in how we must conceive of them if they are to be genuine. Some have suggested that, if we were to demand compensation whenever we experienced difficulty in achieving our aims, this would amount to treating our aims as if they were afflictions or handicaps. And to do this would be to conceive of ourselves in a way that is radically different from the way we ordinarily understand ourselves.[25] The conceptions of the good we affirm and pursue consist of judgments about which aims are valuable and worth pursuing, and these judgments are not ones that we normally would choose to be free from. Even if these judgments are not under our voluntary control, we identify with them because we believe them to be sound.

Let us grant, then, that people, at least in the normal case, reflectively endorse their conception of the good. This means that they will not think of it as an affliction that they would be better off not having. Does this show that people would never have a claim for compensation if they lived in an environment that was inhospitable to their conception of the good? Pretty clearly not. This is evident in the simple case presented above. If the state taxes opera houses for the purpose of discouraging the enjoyment of opera, then those whose conception of the good includes the enjoyment of opera will have a complaint. The state's action in this case contributes to an environment in which it is more costly for them to pursue their conception of the good. In demanding compensation, or in demanding a repeal of the tax, those who affirm this conception of the good would not need to view the fact that they hold this particular conception of the good as an affliction that they would be better off not having. They can reasonably, reflectively endorse this conception of the good and still hold the state responsible for making it harder for them to pursue it

---

[24] Some have argued that this is the relevant division—see Cohen.

[25] Scanlon, "Equality of Resources and Equality of Welfare: A Forced Marriage?" *Ethics*, xcvii (October 1986): 111-18.

This point is important for it shows that it is a mistake to think that, if one reflectively endorses one's conception of the good, then one must conceive of oneself as responsible for all the disadvantages one confronts in pursuing it But, of course, the disadvantage mentioned in this example is not one that a person would suffer in a socially just and neutrally justified state Since the tax on the opera houses is intended to discourage the enjoyment of opera, it runs afoul of neutrality of justification The example, then, is not a counterexample to the second premise of the argument from personal responsibility

But now let us suppose that the tax on the opera houses is neutrally justified Let us assume that it is passed not in order to discourage enjoyment of opera, but in order to raise public revenue Would this show that opera lovers could have no legitimate complaint? It would if we hold them responsible for their taste for opera But why should we hold them responsible for every factor that contributes to the expensiveness of enjoying opera? Suppose our opera lovers could demonstrate that there were other ways in which the state could have raised the public revenue that disadvantaged others in their efforts to pursue their conception of the good to a lesser extent than the tax on opera houses disadvantaged them They then could argue that the tax was unfair to them, and that therefore they are now entitled to compensation

I believe this shows that, when deciding whether or not people are responsible for the difficulty they encounter in pursuing their conception of the good, we should ask not whether the state aimed to disadvantage them, but rather whether they have been treated fairly And a natural way to answer this question is to compare their situation with the situation of others under alternative-possible policies But if we adopt this approach, then we will not have identified a basis for treating environmental disadvantages differently from resource disadvantages Responsibility for both kinds of disadvantage will be apportioned according to the comparative-possibilities approach And far from grounding the asymmetry, this would efface it

The second premise in the argument under consideration, therefore, remains suspect The problem with it is not that people are never responsible for the environmental disadvantages they encounter in pursuing their conception of the good, but rather that sometimes they are not Sometimes they encounter environmental disadvantages for which it would be unfair to hold them responsible, even if the disadvantages were not intended to set back their interests If someone deliberately cultivates expensive tastes or deliberately chooses to take up expensive projects, then it may be unfair to ask others to make sacrifices so that he can better realize his conception of the good In this kind of case, it seems appropriate to hold him responsible for his tastes and projects But so far we have found no good reason for holding persons responsible for *all* the environmental disadvantages they confront in pursuing their conceptions of the good And without such a reason, the argument from personal responsibility does not ground the asymmetry

VI MEASURING FAIR OPPORTUNITIES

I conclude that the asymmetry lacks adequate support Since it lacks adequate support, we should not accept it It is, of course, possible that the asymmetry can be grounded by an argument that I have not considered But, in the absence of this as yet unidentified argument, the asymmetry is ad hoc It arbitrarily limits the state's responsibility for non-neutral outcomes

If we reject the asymmetry, as I am proposing we do, this puts resource and environmental disadvantages on a par Both kinds of disadvantages can and do set back people's interests in pursuing their conception of the good If the comparative-possibilities approach to social justice should apply to the former, then it ought to apply to the latter as well As we have seen, this approach does not invest the distinctions between what the state does and what it allows to happen and between what the state aims to do and what it unintentionally brings about with direct moral significance Accordingly, if one rejects the asymmetry, if one adopts the comparative-possibilities approach to social justice, and if one accepts the principle of state neutrality, then one ought to affirm neutrality of effect over neutrality of justification

Neutrality of effect, as I have conceded, can never be perfectly realized The social/institutional environment sustained by the state will always favor some conceptions of the good over others Still, the view has bite Unlike neutrality of justification, it allows that persons may have a complaint if the neutrally justified state acts in such a way (or fails to act in such a way) that they confront an unfair environmental disadvantage in their efforts to pursue their conception of the good The range and force of such complaints will depend on how the notion of an *unfair* environmental disadvantage is understood Up to this point I have not said much about this I have been concerned to argue only that the class of unfair environmental disadvantages is larger than the class of environmental disadvantages that result from the state's non-neutrally justified actions. Put otherwise, I have been concerned to argue that it is not sufficient to defeat a complaint that a person has suffered an unfair environmental disadvantage to point out that the state did not aim to bring about that disadvantage or that the state acted on the basis of neutral reasons. Sometimes, or so I have suggested, the state is responsible for the environmental disadvantages it causes or allows even if it never aims to bring them about.

The root idea behind neutrality of effect is that each person should have a fair opportunity to realize his conception of the good. This opportunity can be diminished either by a lack of resources or by living in a social/institutional environment that is not conducive to the success of one's aims and aspirations. But how can we know when and to what extent a person has, or has had, a *fair* opportunity to realize his preferred conception of the good? If we accept the comparative-possibilities approach to social justice, then we should say that a person has been denied a fair opportunity to pursue his conception of the good to the extent that he confronts disadvantages, whether resource or environmental, that he would not confront under some alternative feasible set of institutions and public policies that did not impose equally grave or worse disadvantages on others.

This is helpful, but it still leaves it unclear how we should measure the seriousness of these disadvantages. Should this be based on persons' subjective assessments of how difficult it is for them to realize their conceptions of the good, or should it be based on some more objective standard? This question is pressing, since proponents of different conceptions of the good likely will have different ideas about what counts as fulfillment of their aims and aspirations and they likely will have different ideas about which resources and which kinds of social/institutional environments would be most valuable in enabling them to achieve these aims and aspirations. Given this pluralism of assessments, it may seem that we have no option but to refer to the subjective assessments of each person (or group of persons) about the seriousness of the obstacles they confront in pursuing their conception of the good. To invoke an objective standard here would involve the state in making judgments about which substantive goods and opportunities for substantive goods are really valuable and this runs counter, or at least seems to run counter, to the spirit of state neutrality.

Nevertheless, even if the neutralist cannot rely on an objective measure of which opportunities for achieving different conceptions of the good are valuable, he can appeal to various rough-and-ready generalizations. For example, he can say that the value of a given opportunity for a person depends on how central that opportunity is to that person's ability to realize his important aims *and* the extent that it is not possible for him to achieve these aims by pursuing other opportunities that are open to him Admittedly, this is not very precise, but in practice it would allow the neutralist to make judgments that some persons face very poor prospects for realizing their conception of the good In the examples I presented above, both the members of the democratically organized worker's cooperative and the members of the traditional religious community would have a strong claim that they faced serious disadvantages in their efforts to pursue their conception of the good

The ability to make these rough-and-ready judgments is all that the proponent of neutrality of effect needs We should not, accordingly, reject the idea of a fair opportunity to pursue a conception of the good as hopelessly vague or politically unworkable To be sure, further work needs to be done to make this idea more precise, but neutrality of effect remains a serviceable ideal, one that has important policy implications for the modern state

VII CONCLUSION THE COSTS OF NEUTRALITY

My objective has not been to argue directly for or against the principle of state neutrality Instead, I have sought to identify some of the assumptions about the state's responsibilities that underlie different understandings of this principle In particular, I have tried to show that one widely accepted understanding of this principle, neutrality of justification, rests on assumptions that are in tension with a widely accepted approach to social justice, the comparative-possibilities approach

Still, while I have not argued directly against the principle of state neutrality, the argument I have presented, if sound, raises the costs of accepting this principle Many of the leading proponents of state neutrality are committed to the comparative-possibilities approach to social justice They embrace both neutrality of justification and strong egalitarianism The thrust of my argument is that this is not a stable position Neutrality of justification rests on assumptions about the state's responsibility for outcomes that conflict with assumptions about the state's responsibility for outcomes that underlie strong egalitarianism The asymmetry between resource and environmental disadvantages removes the conflict, but, as I have argued, efforts to provide it with an adequate defense have not succeeded

Given this (and assuming a successful defense of the asymmetry is not forthcoming), proponents of state neutrality face some hard choices One option open to them would be to reject strong egalitarianism The case for moving from neutrality of justification to neutrality of effect is premised on the assumption that strong egalitarianism is correct Granting this assumption, a strong case can be made for holding the state responsible for NNO(3), NNO(4), NNO(5), and NNO(6) kinds of outcomes as well as NNO(1) and NNO(2) kinds of outcomes But, if one rejects this assumption, then one would not confront the problem of grounding the asymmetry since one would not have accepted the comparative-possibilities approach to social justice in the first place One could, accordingly, affirm neutrality of justification over neutrality of effect without running into the problems this paper has sought to expose This, then, remains a viable option, but it brings a cost Neutralists can hold on to neutrality of justification, but only if they are prepared to give up strong egalitarianism [26]

A second option would be to hold on to strong egalitarianism and move from neutrality of justification to neutrality of effect This also would bring real costs It would require the state to take a much more active role in intervening to ensure that different groups have fair opportunities to realize their conception of the good than many would be comfortable with The worry that neutrality of effect makes the state unduly intrusive is a genuine worry, but it is one that must be accepted if neutralists are to combine their commitment to neutrality with strong egalitarianism

Neither set of costs is decisive There are neutralists who reject strong egalitarianism, and there are neutralists who embrace neutrality of effect But the most influential and most fully worked out defenses of state neutrality combine neutrality of justification with the comparative-possibilities approach to social justice Coming to see that this is not a stable position is important for assessing the overall case for neutralism For those who remain skeptical of the principle of state neutrality, then, the issue of what the state is and is not responsible for may prove to be an important one for calling into question its plausibility

STEVEN WALL

Kansas State University

---

[26] This, for example, is the position defended by Robert Nozick in *Anarchy, State, and Utopia* (New York Basic Books, 1974)

KNOWING HOW[*]

Many philosophers believe that there is a fundamental distinction between knowing that something is the case and knowing how to do something. According to Gilbert Ryle, to whom the insight is credited, knowledge-how is an ability, which is in turn a complex of dispositions. Knowledge-that, on the other hand, is not an ability, or anything similar. Rather, knowledge-that is a relation between a thinker and a true proposition.

Although few philosophers now share Ryle's general philosophical outlook, his view that knowledge-how is fundamentally different from knowledge-that is widely accepted, so much so that arguments for it are rarely presented, even in the works of those philosophers who crucially rely upon it. For example, Hilary Putnam[1] characterizes the central moral of his work on meaning and understanding in the following terms: "knowing the meaning of the word 'gold' or of the word 'elm' is not a matter of knowing that at all, but a matter of knowing how" (*ibid.*, p. xvi). Yet we are unaware of any passage in which Putnam argues for the distinction. Indeed, even Ryle's positive view that knowledge-how is an ability is widely assumed and crucially exploited in many areas of philosophy outside epistemology. For example, according to David Lewis,[2] knowing what an experience is like amounts to being able to remember, imagine, and recognize the experience. Possession of such abilities, Lewis writes, "isn't knowing that. It's knowing how" (*ibid.*, p. 516). Indeed, according to Lewis, "Know how is ability" (*ibid.*). Similarly, in the philosophy of language, semantic competence is, according to Michael Devitt,[3] "an ability or a skill: a piece of knowledge-how not knowledge-that" (*ibid.*, p. 52, cf. also pp. 23-28).

We contest the thesis that there is a fundamental distinction between knowledge-how and knowledge-that. Knowledge-how is simply a species of knowledge-that. In section I, we discuss Ryle's central argument against the thesis that knowledge-how is a species of knowl-

---

[*] Thanks to Ned Block, Chris Collins, Carl Ginet, Delia Graff, Jeff King, Ernie Lepore, Peter Ludlow, Jim Pryor, Robert van Rooy, Ian Rumfitt, Stephen Schiffer, Paul Snowdon, and Jonathan Vogel for helpful discussion. Thanks also to audiences at Columbia University, Oxford University, Université de Paris/Sorbonne, University College/Dublin, and the University of California/Davis.

[1] Introduction to Andrew Pessin and Sanford Goldberg, eds., *The Twin Earth Chronicles* (Armonk, NY: Sharpe, 1996), pp. xv-xxii.

[2] "What Experience Teaches," in William G. Lycan, ed., *Mind and Cognition: A Reader* (Cambridge: Blackwell, 1990), pp. 499-519.

[3] *Coming to Our Senses* (New York: Cambridge, 1996).

0022-362X/01/9808/411-44

edge-that, as well as Ryle's positive account of knowledge-how in terms of abilities. In section II, we present and defend our positive account of knowledge-how, according to which it is a species of knowledge-that. In section III, we consider and respond to some objections to our view. We conclude by briefly applying our discussion to two uses of the alleged distinction between knowing-that and knowing-how outside epistemology.

I. RYLE ON KNOWLEDGE-HOW

Ryle has two extended discussions of the relation between knowledge-how and knowledge-that.[4] Both have as their main focus the rejection of what Ryle took to be the "prevailing doctrine" of the relation between knowledge-that and knowledge-how, which he took to follow from what he called "the intellectualist legend." This doctrine is the thesis that knowledge-how is a species of knowledge-that. In addition, Ryle presents his own positive view of knowledge-how, according to which, contra the "intellectualist legend," it is not a species of knowledge-that. We begin our discussion by considering Ryle's arguments against what he took to be the prevailing doctrine. Then we turn to his positive account of knowledge-how.

By his own admission, Ryle really had only one argument against the thesis that knowing-how is a species of knowing-that, of which his other arguments were applications. As he writes:

> I rely largely on variations of one argument. I argue that the prevailing doctrine leads to vicious regresses, and these in two directions. (1) If the intelligence exhibited in any act, practical or theoretical, is to be credited to the occurrence of some ulterior act of intelligently considering regulative propositions, no intelligent act, practical or otherwise, could ever begin.... (2) If a deed, to be intelligent, has to be guided by the consideration of a regulative proposition, the gap between that consideration and the practical application of the regulation has to be bridged by some go-between process which cannot by the presupposed definition itself be an exercise of intelligence and cannot, by definition, be the resultant deed (Ryle I, p. 213).

Similarly, and perhaps more clearly, in his later discussion of knowledge-how and knowledge-that:

> The crucial objection to the intellectualist legend is this. The consideration of propositions is itself an operation the execution of which can be more or less intelligent, less or more stupid. But if, for any operation to be intelligently executed, a prior theoretical operation had first to be

---

[4] "Knowing How and Knowing That," in his *Collected Papers, Volume 2* (New York: Barnes and Nobles, 1971), pp. 212-25 [hereafter Ryle I], and *The Concept of Mind* (Chicago: University Press, 1949) [hereafter Ryle II], chapter 2.

> performed and performed intelligently, it would be a logical impossibility for anyone ever to break into the circle (Ryle II, p 30)

It is therefore quite clear that Ryle took his central arguments against the thesis that knowledge-how is a species of knowledge-that to hinge on an accusation of vicious regress Furthermore, Ryle's "vicious regress" argument still carries much weight even in prominent contemporary discussions of the relation between knowledge-how and knowledge-that For example, Edward Craig[5] writes in a recent book

> Hasn't intellectualism been refuted? Ryle, one must admit, gave strong reason for thinking that, when taken in strict generality it must be false it leads to infinite regress I accept the argument and its conclusion (*ibid*, p 154)

Ryle's argument has therefore obviously been influential, so it is worthwhile pausing to examine it in detail[6]

Very roughly, Ryle's argument against the thesis that knowledge-how is a species of knowledge-that is supposed to work as follows if knowledge-how were a species of knowledge-that, then, to engage in any action, one would have to contemplate a proposition But, the contemplation of a proposition is itself an action, which presumably would itself have to be accompanied by a distinct contemplation of a proposition If the thesis that knowledge-how is a species of knowledge-that required each manifestation of knowledge-how to be accompanied by a distinct action of contemplating a proposition, which was itself a manifestation of knowledge-how, then no knowledge-how could ever be manifested

Ryle's argument has two premises

(1) If one *F*s, one employs knowledge how to *F*
(2) If one employs knowledge that *p*, one contemplates the proposition that *p*

Let us see how these two premises operate together to produce a difficulty for the thesis that knowledge-how is a species of knowledge-that

---

[5] *Knowledge and the State of Nature* (New York Oxford, 1990)

[6] Ryle's discussion is standardly taken to show that knowing-how is not a species of knowing-that But his "intellectualist legend" might involve special objectionable features beyond the thesis that knowing-how is a species of knowing-that, and he may have intended his argument also to tell against these additional features Be that as it may, these passages are standardly taken to demonstrate that knowing-how is not a species of knowing-that Our purpose in what follows is to investigate whether this argument establishes the conclusion it is usually taken to establish, and not whether it is successful against the opponent Ryle intended Thanks to Paul Snowdon for discussion here

If knowledge-how is a species of knowledge-that, the content of knowledge how to F is, for some ϕ, the proposition that $\phi(F)$ So, the assumption for reductio is

RA knowledge how to F is knowledge that $\phi(F)$

Furthermore, let '$C(p)$' denote the act of contemplating the proposition that p Suppose that Hannah Fs By premise (1), Hannah employs the knowledge how to F By RA, Hannah employs the knowledge that $\phi(F)$ So, by premise (2), Hannah $C(\phi(F))$s Since $C(\phi(F))$ is an act, we can reapply premise (1), to obtain the conclusion that Hannah knows how to $C(\phi(F))$ By RA, it then follows that Hannah employs the knowledge that $\phi(C(\phi(F)))$ By premise (2), it follows that Hannah $C(\phi(C(\phi(F))))$s And so on

Ryle's argument is intended to show that, if premise (1) and premise (2) are true, then, if knowledge-how is a species of knowledge-that, doing anything would require contemplating an infinite number of propositions of ever-increasing complexity For it to be sound, however, several additional premises are needed First, it must be the case that the function ϕ, which maps acts to propositions, must map distinct actions onto distinct propositions Second, it must be the case that $C(p)$ is a distinct action from $C(\phi(C(p)))$, which is a distinct action from $C(\phi(C(\phi(C(p)))))$, and so on We shall not challenge these additional premises in this section

It is clear from Ryle's own description of his argument, and his reference to "prior theoretical operations," that he thinks that his argument takes the form of a vicious regress But notice that Ryle's argument need not take this form For it to have force, it is not necessary to strengthen premise (2) into the stronger and less well-motivated claim that employment of knowledge-that requires a *prior* action of contemplating a proposition It is simply prima facie implausible that, to engage in an action, it is necessary to contemplate an infinite number of distinct propositions, which, if propositions have structure, would presumably be of ever-increasing complexity If the thesis that knowledge-how is a species of knowledge-that has this consequence, it is surely false

Let us evaluate Ryle's central two premises Take premise (1) first If someone Fs, must they really employ knowledge how to F? This premise is false for many values of 'F' For example, if we instantiate premise (1) to Hannah's action of digesting food, we obtain

(1) If Hannah digests food, she knows how to digest food

But (1) is clearly false Digesting food is not the sort of action that one knows how to do Similarly, if Hannah wins a fair lottery, she still does not know how to win the lottery, since it was by sheer chance that she did so. So, for many values of '*F*', premise (1) is false.

For premise (1) to be true, the range of actions must be sufficiently restricted. Indeed, Ryle hints as much, when he speaks, in the above quotation, of "operations [that are] intelligently executed." Digesting food is not the sort of operation that is executed with intelligence. Similarly, Hannah's winning the lottery was not intelligently performed. Premise (1) is true only when the range of actions is restricted to *intentional actions*. Digesting food is not something done intentionally, and that is why it is not a manifestation of knowledge-how. Similarly, Hannah did not intentionally win the lottery, although she doubtless hoped to win it. So, for premise (1) to be true, the range of actions under consideration must be restricted to intentional actions, or perhaps even a proper subset thereof.

Let us turn to premise (2). As Carl Ginet[7] has pointed out in a neglected brief defense of the thesis that knowledge-how is a species of knowledge-that, on a natural construal of 'contemplation', it is simply false that manifestations of knowledge-that must be accompanied by distinct actions of contemplating propositions. As Ginet writes:

> I exercise (or manifest) my knowledge *that* one can get the door open by turning the knob and pushing it (as well as my knowledge *that* there is a door there) by performing that operation quite automatically as I leave the room, and I may do this, of course, without formulating (in my mind or out loud) that proposition or any other relevant proposition (*op. cit.*, p. 7).

What Ginet's point brings out is that employments of knowledge-that are often unaccompanied by distinct acts of contemplating propositions. So, premise (2) seems straightforwardly false.

There is, however, a way of rescuing premise (2) from Ginet's objection. He clearly construes 'contemplating a proposition' as referring to an intentional act of contemplating a proposition, which is one natural reading of this phrase. If 'contemplating a proposition' is construed in its intentional action sense, then premise (2) is false. But we can rescue premise (2) from Ginet's objection by denying that

---

[7] *Knowledge, Perception, and Memory* (Boston: Reidel, 1975). Two other important defenses of the thesis that knowledge-how is a species of knowledge-that are to be found in David Brown, "Knowing How and Knowing That, What," in O. P. Wood and George Pitcher, eds., *Ryle* (London: Macmillan, 1970), pp. 213-48, and Jaakko Hintikka, "Different Constructions in Terms of the Basic Epistemological Verbs," in his *The Intentions of Intentionality and Other New Models for Modalities* (Boston: Reidel, 1975), pp. 1-25 (cf. pp. 11-14), though Hintikka is less explicit about whether all cases of knowledge-how can be so characterized.

'contemplating a proposition' should be taken in its intentional action sense in premise (2) Perhaps there is a sense of 'contemplating a proposition' in which it refers to an action that is no more intentional than is the action of digesting food Or perhaps it can also be construed as denoting an action merely in some deflationary sense of 'action' If 'contemplating a proposition' is taken in such a sense, then premise (2) can be salvaged after all

As we have seen, however, premise (1) is plausible only if it is restricted to intentional actions If 'contemplates the proposition that *p*' in premise (2) does not refer to an intentional action, then it is not an appropriate substitution instance for '*F*' in premise (1) on its true reading If so, Ryle's argument does not get off the ground There is no uniform reading of the two premises in Ryle's argument on which both are true, the argument is unsound It therefore fails to establish any difficulty for the thesis that knowledge-how is a species of knowledge-that

Let us turn from Ryle's arguments against the thesis that knowledge-how is a species of knowledge-that to his positive account of knowledge-how According to Ryle, an ascription of the form '*x* knows how to *F*' merely ascribes to *x* the ability to *F* It is simply false, however, that ascriptions of knowledge-how ascribe abilities As Ginet and others have pointed out, ascriptions of knowledge-how do not even entail ascriptions of the corresponding abilities For example, a ski instructor may know how to perform a certain complex stunt, without being able to perform it herself[8] Similarly, a master pianist who loses both of her arms in a tragic car accident still knows how to play the piano But she has lost her ability to do so[9] It follows that Ryle's own positive account of knowledge-how is demonstrably false

As we have seen in this section, Ryle's central argument against the thesis that knowledge-how is a species of knowledge-that fails Furthermore, his own positive account of knowledge-how is incorrect But Ryle's positive account of knowledge-how is not the only analysis of knowledge-how according to which it is not a species of knowledge-that For example, David Carr[10] has argued that knowledge-how is a relation between agents and actions, rather than agents and propositions Thus, according to Carr, knowledge-how is a fundamentally different relation from knowledge-that For instance, according to his analysis, sentences such as

(2) Hannah knows how to ride a bicycle

---

[8] Thanks to Jeff King for the example

[9] Cf also Paul Ziff, *Epistemic Analysis* (Boston Reidel, 1984), p 71

[10] "The Logic of Knowing How and Ability," *Mind*, LXXXVIII (1979) 394-409, and "Knowledge in Practice," *American Philosophical Quarterly*, XVIII (1981) 53-61

contain action descriptions, rather than sentential complements as the complement of 'know how' The grammatical objects of ascriptions of knowledge-how, unlike ascriptions of knowledge-that, are therefore nonsentential, the former do not ascribe propositional knowledge Nevertheless, Carr is quite clear that ascriptions of knowledge-how also do not ascribe abilities to agents It is this sort of more sophisticated account of the relation between knowledge-how and knowledge-that which we assume should actually underwrite the current consensus that knowledge-how is a fundamentally different relation from knowledge-that In the next section, we present our own positive account of knowledge-how, according to which it is a species of knowledge-that In the course of it, we undermine even these sorts of more sophisticated accounts of the relation between knowledge-how and knowledge-that

II A POSITIVE ACCOUNT OF KNOWLEDGE-HOW

According to the more sophisticated account of the distinction between knowledge-how and knowledge-that found in contemporary defenders of Ryle's distinction, sentences such as (2) have a distinct syntactic structure from sentences such as (3)

(2) Hannah knows how to ride a bicycle
(3) Hannah knows that penguins waddle

William Bechtel and Adele Abrahamsen[11] write, in describing what they call the "linguistic distinction" between constructions such as (2) and (3) "In general, the expression 'knowing that' requires completion by a proposition, whereas the expression 'knowing how' is completed by an infinitive (e g 'to ride') specifying an activity" (*ibid*, p 151) On this view, in a sentence such as (2), 'knows how' forms a constituent, which takes as a complement the expression 'to ride a bicycle', which is a description of an action 'Know' has no clausal complement in (2) In (3), on the other hand, 'that penguins waddle' is the clausal complement of 'knows', and denotes a proposition, which is the object of the knowledge relation

Such accounts of the syntactic structure of sentences like (2), however, are inconsistent with what is said about such structures in recent syntactic theory Although syntactic frameworks have undergone much change since the early 1970s, none of it has affected the basic analysis of sentences such as (2), and its syntactic counterparts, such as

---

[11] *Connectionism and the Mind* (Cambridge Blackwell, 1991) Assuming propositions to be nonlinguistic entities, as we do, there is an obvious use-mention difficulty in Bechtel and Abrahamsen's claim Nonetheless, their intent is clear

(4) (a) Hannah knows where to find a nickel
(b) Hannah knows whom to call for help in a fire
(c) Hannah knows which prize to look for
(d) Hannah knows why to vote for Gore[12]

(2) and (4a-d) have two, and only two syntactic features which distinguish them from sentences such as (3) First, (2) and (4a-d) contain what are called embedded questions, whereas (3) obviously does not contain an embedded question Second, (2) and (4a-d) contain *untensed* clauses, whereas (3) contains a *tensed* clause There is, of course, no conceptual connection between these two syntactic features Embedded questions may easily occur in tensed clauses, as in

(5) (a) Hannah knows how Bill rides a bicycle
(b) Hannah knows whom Bill called for help yesterday
(c) Hannah knows which prize Mary was looking for
(d) Hannah knows why Gore is the best candidate

It is also worth mentioning that embedded 'how' questions with untensed clauses can occur with all sorts of verbs, and are not restricted to co-occurring with 'know' For example

(6) (a) Hannah learned how to ride a bicycle
(b) Hannah recalled how to ride a bicycle
(c) Hannah asked how to ride a bicycle
(d) Hannah wonders how to ride a bicycle
(e) Hannah is certain about how to ride a bicycle
(f) Hannah indicated how to ride a bicycle
(g) Hannah saw how to ride a bicycle

From the perspective of their syntactic structure, there are no relevant differences between sentences such (2) and (6a-g) This suggests that it is incorrect to take 'know how' as a constituent in sentences such as (2)[13]

---

[12] We have not included, on this list, sentences involving 'whether', such as 'Hannah knows whether to eat meat on Sundays' The reason is that 'whether' is sufficiently syntactically distinct from 'how', 'who', 'where', 'why', and 'which' so that incorporation of it would distract from the discussion to follow Unlike these words, 'whether' is commonly thought not to undergo movement, but rather to be generated in the position it appears to have on the surface Furthermore, constructions such as 'Hannah knows whether or not to eat meat on Sundays' are grammatical, while 'John knows who or not to call' or 'John knows how or not to ride a bicycle' are not These syntactic differences do not, however, affect the standard semantics for embedded questions, 'Hannah knows whether to eat meat on Sundays' receives the same semantic treatment as (4a-d)

[13] Ryle himself noticed this point, mentioning it as a "parallel" between knowing-how and knowing-that (Ryle II, p 28) In the next paragraph, he points to the fact that we never speak of 'believing how' as a "divergence" between knowing-how and

Since the standard philosophical examples of ascriptions of knowledge-how occur with untensed clauses, as in (2), rather than tensed clauses, as in (5), it is the constituent structure of those which we shall first discuss Where brackets signal clausal boundaries, abstracting from the many irrelevant details, the standardly accepted constituent structure of embedded questions with untensed clauses, as in (2) and (4a-d), is as follows

(7) (a) Hannah knows [how PRO to ride a bicycle *t*]
(b) Hannah knows [where PRO to find a nickel *t*]
(c) Hannah knows [whom PRO to call *t* for help in a fire]
(d) Hannah knows [which prize PRO to look for *t*]
(e) Hannah knows [why PRO to vote for Gore *t*]

'PRO' here is a phonologically null pronoun that occurs, according to standard syntactic theory, in the subject position of untensed clauses The occurrences of '*t*' in (7a-e) are the *traces* of movement of the phrases 'how', 'where', 'whom', 'which person', and 'why', respectively These traces occur at the site from which the phrases have been moved [14]

The constituent structures of sentences involving embedded questions in tensed clauses, such as (5a-d), differ only in that they contain overt noun phrases where the phonologically null pronoun 'PRO' occurs in their untensed counterparts But an embedded question in a tensed clause, such as

(5a) Hannah knows how Bill rides a bicycle

seems clearly to attribute propositional knowledge to Hannah As we have seen, from the perspective of syntactic theory, there is no difference between (2) and (5a) which would lead us to think that (2) ascribes nonpropositional knowledge, whereas (5a) ascribes propositional knowledge The supposed difference has no basis in structure There are indeed interesting distinctions between embedded questions in tensed clauses and those in untensed clauses relating to the occurrence of 'PRO' As we shall see, these distinctions on their own explain the intuitions that might lead one incorrectly to the thesis that (2) ascribes nonpropositional knowledge, whereas (5a) ascribes propositional knowledge

---

knowing-that But the reason we never speak of 'believing how' is simply that 'believes' does not take embedded questions (cf the oddity of 'John believes why to vote for Gore') Though we may speak of 'believing what John says', the 'what' here is the relative pronoun 'what' (as in 'whatever'), what we have in this construction is a noun phrase complement of 'believe', rather than an embedded question

[14] When it is unnecessary, we shall omit reference to the traces of movement in the discussion to follow

Of course, the standardly accepted constituent structure of (3) is

(8) Hannah knows [that penguins waddle]

So, in both constructions such as (2) and constructions, such as (3), 'know' takes a sentential complement The syntactic difference between sentences such as (2) and sentences such as (3) is just that the former contain embedded questions with untensed clauses

Let us now turn from the standardly accepted syntax of constructions such as (2), to the standardly accepted semantics There are a variety of different classical treatments of the semantics of embedded questions in the literature But we shall, with only minimal commentary, use Lauri Karttunen's[15] classic 1977 account in presenting our account of embedded questions such as (2) Our account does not rely on any of the special features of Karttunen's theory It is easily translatable into other major contemporary frameworks without altering any of the substance of our claim that ascriptions of knowledge-how are ascriptions of propositional knowledge [16]

According to Karttunen's theory, an embedded question denotes the set of its true answers To illustrate Karttunen's semantics, consider the following simple construction involving an embedded question

(9) Hannah knows whom Bill likes

On Karttunen's analysis, the embedded question 'whom Bill likes' denotes the set of true propositions expressed by sentences of the form 'Bill likes *x*' (9) is true if and only if, for each proposition *p* in that set, Hannah knows that *p* [17]

---

[15] "Syntax and Semantics of Questions," *Linguistics and Philosophy*, i (1977) 3-44 As Jeroen Groenendijk and Martin Stokhof have written in a recent comprehensive survey of the literature on the topic—"Questions," in Van Benthem and Ter Meulen, eds , *Handbook of Logic and Language* (Cambridge MIT, 1997), pp 1055-124, Karttunen's approach "is probably the most influential analysis in the semantics literature to date" (p 1105)

[16] Two other now classical theories of questions are C L Hamblin, "Questions in Montague English," *Foundations of Language*, x (1973) 41-53, and James Higginbotham and Robert May, "Questions, Quantifiers, and Crossing," *Linguistic Review*, i (1981) 41-79 But both these theories are more concerned with unembedded questions Higginbotham and May's discussion also mainly involves topics irrelevant to our central concerns, such as the difficulties involved in interpreting questions containing multiple wh-phrases

[17] The point that ascriptions of knowledge-who, knowledge-what, knowledge-when, and knowledge-where are fundamentally ascriptions of propositional knowledge of course antecedes Karttunen by many years in the philosophy literature For one quite early reference, see Hintikka's *Knowledge and Belief An Introduction to the Logic of the Two Notions* (Ithaca Cornell, 1962), pp 131-32

Informally, this is how Karttunen's semantics works The question-embedding verb 'know', which we may symbolize as 'know_q', takes two arguments, a person, and

For (9) to be true, must Hannah know all the propositions in the set denoted by the embedded question? In certain contexts, (9) might be felicitously uttered when all are aware that Hannah stands in the knowledge-that relation only to a few propositions in the set of true propositions expressed by sentences of the form 'Bill likes *x*' Nonetheless, it is in fact standard to assume that, for (9) to be true, Hannah must know all of the propositions in the denotation of the embedded question This prediction can be seen to be correct, once the account is adjusted to accommodate the effects of extralinguistic context on constructions involving embedded questions For a particular use of (9) to be true, Hannah must know all propositions in the denotation of the embedded question *relative to that context of use* For example, a particular use of (9) may mean that Hannah knows whom Bill likes in Bill's math class Relative to such a context, the denotation of the embedded question is the set of true propositions expressed by sentences of the form 'Bill likes *x* & *x* is in Bill's class' In what follows, we shall take the set of propositions denoted by the embedded question to so depend on extralinguistic context [18] In many contexts, the set may contain very few members [19]

Up until this point, we have been discussing embedded questions involving what linguists call *arguments* rather than *adjuncts* In (9), the word 'whom' originates in argument position, as the complement of the transitive verb 'likes' In contrast, words like 'why' and 'how' are what linguists call adjuncts rather than arguments, and so do not originate in argument position [20] But this difference is completely irrelevant to the semantics of embedded questions Karttunen's se-

---

the set of propositions, which is the denotation of the embedded question (the set of true answers to that question) The lexical meaning of the question-embedding verb 'know' is such that *x* stands in the relation expressed by 'know$_q$' to a set of propositions if and only if *x* knows all of the propositions in that set Thus, on Karttunen's view, while there is strictly speaking an ambiguity (or, more accurately, polysemy) between question-embedding uses of 'know' and normal clausal-complement uses of 'know', the former are analyzed in terms of the latter

[18] This point is standard in discussions of embedded questions, cf Groenendijk and Stokhof, "Semantic Analysis of Wh-complements," *Linguistics and Philosophy*, v (1982) 175-233, especially pp 180-81

[19] What if there is no one Bill likes? The worry here is that sentence (9), 'Hannah knows whom Bill likes', will be trivially true relative to such a context But (9) semantically presupposes that Bill likes someone If this presupposition is not satisfied, no proposition is expressed Karttunen and Groenendijk and Stokhof all assume that what appears to be a classical case of presupposition in fact is an implicature But their central reasons seem just to stem from a general hostility to semantic presupposition (cf Groenendijk and Stokhof, p 184)

[20] For example, adjuncts have slightly different movement properties from arguments, especially with regard to so-called "weak islands" For one classic discussion of this topic, see Luigi Rizzi's *Relativized Minimality* (Cambridge MIT, 1990)

mantics is intended to apply equally to embedded questions headed by adjuncts, as in

(10) Hannah knows why Bill votes Republican *t*

On this account, the embedded question 'why Bill votes Republican *t*' denotes the set of true propositions expressed by sentences of the form 'Bill votes Republican for reason *r*' (10) is true if and only if, for all propositions *p* in this set, Hannah knows that *p*

Let us now consider some of the more complicated examples of embedded questions which we have been discussing Of particular relevance are the complications involved in interpreting embedded questions in untensed clauses, as in

(7c) Hannah knows [whom PRO to call *t* for help in a fire]

As we have discussed, 'PRO' is an empty pronominal element which occurs in the subject position of infinitives in English There are two complicating factors in interpreting such constructions The first involves the proper interpretation of 'PRO' The second involves the interpretation of the infinitive There are sizable literatures in linguistics on both of these topics The complications raised below, however, have no bearing on the thesis that knowledge-how is a species of knowledge-that (7b-e) all clearly attribute propositional knowledge to Hannah Furthermore, their cousins

(11) (a) Hannah recalled where PRO to find a nickel *t*
(b) Hannah asked why PRO to vote for Gore *t*
(c) Hannah discovered whom PRO to call *t* for help

all clearly involve relations between persons and propositions It is just that, in all these cases, the occurrence of 'PRO' and the use of the infinitive makes it somewhat tricky to state exactly which propositions are at issue

The first complication involved in the interpretation of embedded questions in untensed clauses involves the interpretive possibilities for 'PRO' When 'PRO' occurs in an untensed clause that is not an embedded question, as in the sentences in (12), it receives its interpretation obligatorily from the subject of the main clause (as is standard, co-indexation with an element to its left represents a relation of referential dependence)

(12) (a) Hannah$_i$ wants PRO$_i$ to win the race
(b) Hannah$_i$ expects PRO$_i$ to cook pasta tonight

Because of this limitation on the interpretation of 'PRO' in such contexts, they are standardly called contexts of *obligatory control*

In embedded questions in untensed clauses, the favored reading of 'PRO' links it to the subject of the main clause, as in the examples in (12) Syntacticians have known for some time, however, that embedded questions in untensed clauses are not contexts of obligatory control In contexts of obligatory control, 'PRO' cannot be interpreted in the so-called 'PRO-arbitrary' manner, where it is roughly equivalent to 'one' This is demonstrated by the ungrammaticality of

(13) * Hannah wants PRO to behave oneself

But embedded questions in untensed clauses seem to allow 'PRO' to be interpreted in this manner, as the following example demonstrates[21]

(14) John asked how PRO to behave oneself

Here is why, in (14), 'PRO' is interpreted as 'one' 'Oneself' is an anaphor in the technical syntactic sense Anaphors require their antecedent to be within the same minimal clause The occurrence of 'oneself' could be licensed only if 'PRO' is interpreted as 'one'

Of course, (14) is less natural than

(15) John asked how to behave himself

But (14) is nonetheless far more acceptable than (12) Examples with a similar level of acceptability are

(16) (a) John knows how PRO to shoot oneself
 (b) John knows why PRO to shoot oneself
 (c) John knows how PRO to kill oneself painlessly
 (d) Bill wondered how PRO to shave oneself without hot water

In all of these cases, 'PRO' must be interpreted as 'one' The conclusion in the linguistics literature from such data is that occurrences of 'PRO' in embedded questions in untensed clauses can be interpreted either as expressions anaphoric on the subject of the main clause, or, albeit less naturally, as 'one'

The second complication concerns the interpretation of the infinitives in untensed embedded questions The difficulty in interpreting infinitives in such constructions is that they do not have any obvious tense Indeed, on one very standard use of infinitives, they have interpretations with some sort of deontic modal force, as in

---

[21] The example is from Noam Chomsky's "On Binding," *Linguistic Inquiry*, xi, 1 (1980), but this application of it is due to Maria Rita Manzini's "On Control and Control Theory," *Linguistic Inquiry*, xiv (1983) 421-46

(17) (a) Hannah is the person to call in case of danger
(b) The screwdriver to use is a Phillips
(c) A person to call when in need of assistance with moving is someone with no back trouble

For example, (17a) expresses something similar to what is expressed by the sentence, 'Hannah is the person one ought to call in case of danger'

But infinitives also have readings on which they have a different kind of modal force For example, consider

(18) John asked where to board the plane

The infinitive in (18) has a natural construal not in terms of deontic modality According to this reading, John did not ask where he ought to board the plane Rather, John asked where he *could* board the plane That is, John wants as a response a proposition whose informational content is something like that expressed by '*w* is a place for John to board the plane' [22]

So, infinitives appear to have at least two different kinds of readings On the first reading, they express deontic modality In this case, a use of 'to *F*' expresses something like 'ought to *F*' On the second reading, they express some kind of possibility On this reading, a use of 'to *F*' expresses something like 'can *F*' These are the two readings which seem available for infinitives in embedded questions, and hence are the ones relevant for our purposes here

Let us sum up the discussion of the two complications involved in the interpretation of embedded questions in untensed clauses Occurrences of 'PRO' in such constructions have two interpretive possibilities According to the first, 'PRO' receives its interpretation from the subject of the main clause According to the second, 'PRO' means something like 'one' An infinitive in such a construction also has two relevant interpretations According to the first, the infinitive has ought'-like force According to the second, the infinitive expresses 'can'-like force So, in general, we would expect four interpretive possibilities for sentences containing embedded questions in untensed clauses

Consider now

(19) Hannah knows how PRO to ride a bicycle

In such an example, we should expect the embedded question to have four interpretive possibilities, corresponding to (20a-d)

---

[22] Cf Brown, p 240

(20) (a) Hannah knows how she ought to ride a bicycle
(b) Hannah knows how one ought to ride a bicycle
(c) Hannah knows how she could ride a bicycle
(d) Hannah knows how one could ride a bicycle

(19) and its cousins certainly have these interpretive possibilities[23] The interpretations given in (20a) and (20b) quite obviously seem to attribute some kind of propositional knowledge to Hannah, so they are not the interpretations underlying the thesis that knowledge-how is not a species of knowledge-that It is rather interpretations such as (20c) and (20d) that seem to be at issue in philosophical discussions of knowledge-how So, let us see what is predicted by applying Karttunen's semantics to (19), interpreted in one of these latter two ways

Consider (20c), the paradigm reading of (19) on which we shall focus in the rest of this discussion On this reading of (19), 'PRO' receives its interpretation from 'Hannah', and the infinitive has some kind of nonnormative modal force Interpreted in this way, Karttunen's semantics predicts that (19) is true if and only if, for all propositions *p* expressed by sentences of the form '*w* is a way for Hannah to ride a bicycle', Hannah knows that *p*[24]

Now, clearly Hannah need not know all such propositions for (19) to be true on this reading It is more natural to construe (19) (on this reading) as true if and only if Hannah knows *some* proposition of the relevant form, that is, for *some* way *w*, Hannah knows that *w* is a way for Hannah to ride a bicycle[25] But this is not a semantic difference between constructions like (19) and other constructions involving other kinds of embedded questions Constructions involving embedded questions are generally ambiguous between what Jeroen Groenendijk and Martin Stokhof[26] have called the "mention-all" and "mention-some" readings For example, the sentences in (21a-b)), in addition to the readings we have been discussing above (Groenendijk and Stokhof's "mention-all" readings), allow for the readings in (22a-b)

(21) (a) John knows where to find an Italian newspaper
(b) John knows who has a light

---

[23] For example, consider the sentence 'Hannah knows how PRO to ride a bicycle in New York City (namely carefully)' Here, the infinitive has a clear ought-like modal force

[24] This is also the interpretation suggested by the brief but prescient discussion of the 'knowing the way' sense of ascriptions of knowledge-how in Hintikka, *The Intentions of Intentionality*, pp 11-14

[25] Cf Brown, p 240

[26] *Studies on the Semantics of Questions and the Pragmatics of Answers* (University of Amsterdam Dissertation, 1984), chapter 6

(22) (a) John knows, of some place *p*, that *p* is a place to find an Italian newspaper
(b) John knows, of some person *x*, that *x* has a light

Similarly, although embedded questions involving 'know how' often favor the "mention-some" reading, they are nevertheless ambiguous between these two readings For example, the examples in (23) have natural "mention-all" readings, where the subject of the sentence is said to know something about all contextually relevant ways

(23) (a) The warden of the prison knows how to escape from it
(b) The expert pitching coach knows how to pitch to a dangerous switch-hitter

The fact that (19) is more often used with the "mention-some" reading in mind is due to the distinctive communicative purpose of the relevant class of uses of (19)

When we inform our audience that *x* knows how to *F*, using it in the sense of (20c), it is irrelevant to our communicative purpose that our audience come to know, of every way which is a way for *x* to *F*, that *x* knows that *w* is a way for *x* to *F* It is sufficient for our communicative purpose to inform our audience that *x* knows of one such way that it is a way for *x* to *F* In cases such as (9), by contrast, our communicative purpose involves informing our audience that Hannah knows, of every relevant person whom Bill likes, that Bill likes that person This difference in communicative purpose is the reason why speakers more often use '*x* knows how to *F*' with the "mention-some" reading in mind than the "mention-all" reading

In what follows, we restrict our attention to the "mention-some" reading of constructions such as (19) Relative to a context in which (19) is interpreted as (20c), (19) is true if and only if, for some contextually relevant way *w* which is a way for Hannah to ride a bicycle, Hannah knows that *w* is a way for her to ride a bicycle[27]

Thus, to say that someone knows how to *F* is always to ascribe to them knowledge-that To complete our account, however, we need to say which of the standard theories of propositional attitudes we are adopting Essentially, there are three standard theories of the semantics of propositional-attitude ascriptions According to the first two, sentences such as

---

[27] Incorporating this insight into Karttunen's framework requires making a distinction between two question-embedding verbs, 'know$_V$' and 'know$_I$' Someone stands in the relation expressed by 'know$_I$' to a set of propositions if and only if she knows at least one of the propositions in that set For a discussion of how to incorporate the 'mention-some' readings of embedded questions into a semantic framework that is a development of Karttunen's, see Groenendijk and Stokhof, *ibid*, pp 528ff

(24) (a) Hannah believes that penguins waddle
(b) Hannah knows that she can walk

relate persons and propositions. The two theories differ, however, on the nature of propositions. According to the first theory, the contemporary Russellian theory, propositions are ordered sequences of properties and objects. According to the second theory, the Fregean theory, propositions contain modes of presentations of properties and objects, rather than the properties and objects themselves. Finally, according to the third standard theory, verbs such as 'believes' and 'knows' express three-place relations between persons, Russellian propositions, and ways of thinking of Russellian propositions.

Our view can be stated in any of these three frameworks. For clarity's sake, however, we shall take propositions to be Russellian, as in the first and third of these theories. The propositions that concern us will contain *ways of engaging in actions*. To be more precise, we shall take ways to be properties of token events. Ways are the elements of the domain of quantified expressions such as 'however', as in

(25) However Douglas passes the ball, it results in a basket

We believe that any successful account of natural language must postulate entities such as ways. But we shall not have much more of substance to say about the metaphysics of ways in this paper.

So, according to our official account, on the relevant understanding of (19), its truth requires Hannah to stand in the knowledge-that relation to a Russellian proposition containing a way of riding a bicycle (along with other objects and properties). But our account is still incomplete. If a way is really a property, an element of a standard Russellian proposition, then it must be possible for it to be entertained under different modes of presentation. And we have not yet said anything about the modes of presentation under which propositions containing ways may be entertained.

According to some accounts involving modes of presentation, the particular mode of presentation under which a person entertains a Russellian proposition is irrelevant to the truth conditions of a corresponding attitude ascription; it is only of pragmatic relevance. According to other theorists, such as John Perry and Mark Crimmins,[28] context provides a particular mode of presentation under which the proposition is entertained to the truth conditions of a propositional-attitude ascription. In what follows, we shall speak of modes of presentation as being associated with certain linguistic constructions, but remain neutral on the question of whether these modes of presentation have semantic import.

---

[28] "The Prince and the Phonebook: Reporting Puzzling Beliefs," this JOURNAL, LXXXVI, 12 (December 1989): 685-711.

It is relatively straightforward to show that one and the same way can be entertained under distinct modes of presentation. But before we turn to this task, it is useful to review the general strategy for constructing such cases. Suppose that John is looking in a mirror, which he mistakenly believes to be a window. Seeing a man whose pants are on fire, and not recognizing that man as himself, John forms the demonstrative belief that that man is on fire. Intuitively, however, John does not believe that his own pants are on fire. That is, relative to the envisaged context, (26) is true and (27) is false.

(26) John believes that that man has burning pants.
(27) John believes that he himself has burning pants.

Given that 'that man' refers to John, however, the complement clauses of (26) and (27) express the same proposition, namely, the singular proposition containing John. To distinguish between (26) and (27), contemporary advocates of Russellian propositions appeal to different modes of presentation under which that proposition is entertained. In the envisaged context, (26) is associated with a demonstrative mode of presentation (or guise) of the relevant proposition, whereas (27) is associated with a first-personal mode of presentation of that very same proposition.

Here is a parallel case involving ways. Suppose that Hannah does not know how to ride a bicycle. Susan points to John, who is riding a bicycle, and says, 'That is a way for you to ride a bicycle'.[29] Suppose that the way in which John is riding his bicycle is in fact a way for Hannah to ride a bicycle. So, where the demonstrative 'that way' denotes John's way of riding a bicycle, (28) seems true.

(28) Hannah knows that that way is a way for her to ride a bicycle.

Relative to this context, however,

(29) Hannah$_i$ knows [how PRO$_i$ to ride a bicycle]

---

[29] Since ways are properties rather than particulars, such demonstrative reference involves deferred ostension to a property. That is, what Susan actually ostends is an instance of a way of riding a bicycle, and her demonstrative expression refers to a property of which that is an instance. But this does not prevent (25) from involving a de re mode of presentation of the property in question (that is, the way of riding a bicycle). To take a parallel example, suppose Susan, pointing at a brown chairs, utters 'I know that that color sickens John'. Here, Susan's demonstrative reference "exploits the presence of the sample" to refer to the color—John McDowell, *Mind and World* (Cambridge: Harvard, 1994), p. 57. But her utterance is still associated with a de re mode of presentation of the color; indeed, it is a paradigm example of such an association.

seems false This case parallels (26) and (27) Where the demonstrated way is the only contextually salient way of riding a bicycle, (28) and (29) ascribe knowledge of the same proposition to Hannah But this proposition is ascribed under different guises In (28), knowledge of the proposition is ascribed to Hannah under a demonstrative mode of presentation In (29), knowledge of that proposition is ascribed to Hannah under a different mode of presentation, what we call a *practical* mode of presentation [30]

There is a conventional connection between pronouns such as 'he himself' and first-person modes of presentation Similarly, there is a conventional connection between the use of constructions that embed instances of the schema 'how to *F*', and practical modes of presentations of ways Such conventional connections between linguistic constructions and modes of presentation have obvious communicative value in both the case of first-person propositional-attitude ascriptions and the constructions that concern us in this paper Given such a connection, use of the relevant construction provides extra information about how the ascribee thinks about one of the propositional constituents, information that allows the hearer to predict how the ascribee will act in various situations [31]

Giving a nontrivial characterization of the first-person mode of presentation is quite a substantial philosophical task Unfortunately, the same is true of giving a nontrivial characterization of a practical mode of presentation of a way In both cases, however, one can provide an existence proof for such modes of presentation If, as is assumed in much of philosophy of language, there is a sound argument from (26) and (27) to the existence of first-personal guises of propositions, then there is a sound argument from (28) and (29) to the existence of practical guises of propositions

Thinking of a person as oneself entails being disposed to behave in certain ways, or form certain beliefs, given relevant input from that person Similarly, thinking of a place as *here* entails being disposed to behave in certain ways, or form certain beliefs, given relevant input from that place Analogously, thinking of a way under a practical mode of presentation undoubtedly entails the possession of certain complex dispositions It is for this reason that there are intricate connections between knowing-how and dispositional states But acknowledging such connections in no way undermines the thesis that knowing-how is a species of knowing-that For example, such connections are also present in the case of first-person thought But this in no way threatens the thesis that thought about oneself is genuinely propositional It is simply a feature of certain kinds of propositional knowledge that possession of it is related in complex ways to dispositional states Recognizing this fact eliminates the need to postulate a distinctive kind of nonpropositional knowledge

---

[30] If someone entertained a way of riding a bicycle by possessing a complete physiological description of it, that might also give them de re knowledge of that way, though not under a practical mode of presentation Whether or not it does depends upon what acquaintance with properties requires

[31] To say that there are such conventional connections does not preclude them from being of only pragmatic significance As stated above, we are neutral on the issue of whether modes of presentation affect semantic content

So, here is our complete account of knowing-how Suppose modes of presentation are semantically relevant Then (29) is true relative to a context *c* if and only if there is some contextually relevant way *w* such that Hannah stands in the knowledge-that relation to the Russellian proposition that *w* is a way for Hannah to ride a bicycle, and Hannah entertains this proposition under a practical mode of presentation If modes of presentation are not semantically relevant, then the truth of (29) does not require that Hannah entertain the proposition in question under a practical mode of presentation, though a use of (29) pragmatically conveys that she does

We have exploited Karttunen's account of embedded questions in developing our own account of examples such as (19) It is worth mentioning that later semantical frameworks modify Karttunen's analysis in various directions For example, one problematic aspect of Karttunen's framework is that embedded questions are taken to denote sets of propositions, whereas 'that' clauses denote propositions There is some evidence that both embedded questions and 'that' clauses denote entities of the same type, namely, propositions For example, in linguistics, the possibility of conjunction is a standard indication of sameness of semantic type And, as Groenendijk and Stokhof[32] point out, one can conjoin embedded questions and 'that' complements, as in

(30) (a) John knows both that Peter has left for Paris and whether Mary has followed him
(b) Alex told Susan that someone was waiting for her, but not who it was

A similar point holds for embedded questions headed by 'how', as evidenced by the examples in (31)

(31) (a) John knows both how to ride a bicycle and that accidents can happen to anyone

---

[32] "Semantic Analysis of Wh-complements," p 185

(b) John knows both how to twitch his ears and that his mother is sickened by facial tricks
(c) John knows both that his mother hates facial twitches and how to make them

What such examples suggest is that we should take embedded questions, including those involved in the ascription of knowledge-how, to denote propositions, rather than sets of propositions Incorporating this insight only strengthens the parallels between ascriptions of knowledge-that and ascriptions of knowledge-how But it also involves the introduction of technical machinery that would distract from our central points We shall therefore continue to couch our views in Karttunen's framework[33]

If these standard accounts of the syntax and semantics of embedded questions are correct, then ascriptions of knowledge-how simply ascribe knowledge-that to their subjects The propositions that the subjects of ascriptions of knowledge-how are said to know in such ascriptions involve ways of engaging in actions But this does not make them any less propositional

Our view of ascriptions of knowledge-how is very straightforward It is just that the standard linguistic account of the syntax and semantics of embedded questions is correct Furthermore, it should not be radically altered to rescue philosophical views about an allegedly philosophically significant subclass of them Sentences such as (2) have sentential complements Furthermore, a sentence such as (2) is true relative to a context if and only if the subject stands in the knowledge-that relation to some (or every) member of the set of propositions denoted by the embedded 'how' question in that context

We take our view of ascriptions of knowledge-how to be the default position From a linguistic perspective, very little is special about ascriptions of knowledge-how It is hard to motivate singling them out for special treatment from the rest of a family of related constructions Our view of ascriptions of knowledge-how is the analysis reached on full consideration of these constructions by theorists unencumbered by relevant philosophical prejudices[34]

---

[33] On Groenendijk and Stokhof's theory, the semantic value of an embedded question is a propositional concept, a function from indices to propositions But the question embedding verb 'know', on their view, is extensional That is, it operates on the denotation of the propositional concept expressed by its embedded clause, relative to the world of context The denotation of a propositional concept relative to a world of context is a proposition

[34] Like us, Brown (*op cit*) looks to linguistics for enlightenment about the true structure of constructions such as 'Hannah knows how to ride a bicycle' At the time at which Brown was writing, however, little progress had been made on the syntax

Of course, we would like to compare the standard syntactic and semantic accounts of such constructions with the syntactic and semantic accounts of them given by those who follow Ryle in rejecting the thesis that knowledge-how is a species of knowledge-that. Surprisingly, however, none of Ryle's followers has ever given an explicit syntax and semantics for such constructions, much less one which would give them the interpretations they claim such constructions to have. Therefore, such a comparison is impossible. Nonetheless, we recognize the possibility that the enormous amount of attention philosophers have directed to this linguistically rather ordinary construction may be motivated by special features to which linguists have inexplicably been blind. To eliminate this possibility, we must show that our analysis in fact accounts for any such special feature. That is the task of the next section.

III. A DEFENSE

Consider again our paradigm ascription of knowledge-how:

(29) Hannah$_i$ knows [how PRO$_i$ to ride a bicycle].

Abstracting from the possible semantic relevance of modes of presentation, on our analysis, (29) is true relative to context if and only if, for some contextually relevant way *w* for Hannah to ride a bicycle, Hannah knows that *w* is a way for Hannah to ride a bicycle. Our proposal reflects the intuitively felt connection between 'Hannah knows how to ride a bicycle' and 'Hannah knows in what way to ride a bicycle'. In the course of responding to objections to our account, we now show that it explains the special features philosophers have claimed such constructions to have.

We shall now consider a series of worries about our proposal. Here is the first. Even if Hannah knows how to ride a bicycle, there need be no informative sentence of the form 'I ride a bicycle by *F*ing' which she would recognize as true. That is, for (29) to be true, there need be no sentence she understands and accepts containing a purely nonindexical description of a way of riding a bicycle.

Let us consider a parallel case. Consider:

(32) Hannah knows that she is in pain.

Suppose that (32) attributes a first-personal belief to Hannah. It is notoriously difficult to explicate the notion of a first-personal guise of

---

and semantics of embedded questions. Brown nevertheless manages to make an impressive amount of progress on the topic, despite the lack of linguistic analyses available to him.

a proposition As is familiar from the work of Perry,[35] (32) may be true, even if there is no pure nonindexical description 'The *F*' of Hannah, for which she would accept the sentence 'The *F* is in pain' To take an extreme case, Hannah may have amnesia after a tragic accident, and have completely forgotten any purely descriptive uniquely identifying facts about herself In such a case, (32) may still be true, even though there is no purely descriptive sentence of the form 'The *F* is in pain' that Hannah would accept

But we need appeal to nothing so philosophically controversial as first-person thought in order to respond to the objection Suppose that Hannah rides a bicycle in a most peculiar manner John is unable to describe the way in which Hannah rides a bicycle, but he can physically imitate it In trying to convey how Hannah rides a bicycle, he imitates her motions, and says

(33) I know that Hannah rides a bicycle in this way

John's knowledge here is propositional knowledge involving a way John's use of 'this way' refers to this way But, as we said, John is unable to describe in nonindexical involving terms the way in which Hannah rides a bicycle So, completely apart from issues involving knowledge-how, that John possesses propositional knowledge about a way does not entail that he can describe it in nonindexical terms To say that (29) is a propositional-knowledge ascription simply does not imply that the guises of the relevant propositions can be described in nonindexical terms [36]

Some still believe that all mental states can be characterized in nonindexical terms We are not sympathetic to this view But if it is correct, there is a nonindexical description of John's thought about Hannah's way of riding a bicycle that John would accept If such a description can be provided for (33), it can also be provided for (29)

Here is a second worry with our proposal [37] We have appealed, in our account of knowledge-how, to practical modes of presentation One might worry then that we have not succeeded in our aim of establishing that knowledge-how is just knowledge-that, since an analysis of these practical modes of presentation might require appeal to an unreduced notion of knowing-how

This worry misunderstands our present purpose, however We are not engaged in the reductive project of reducing talk of knowledge-how to talk that does not involve knowledge-how Our view is rather that knowledge-how is a species of knowledge-that To establish that one concept is a subspecies of another, it is not necessary to provide a reductive analysis of that concept For example, perhaps the concept expressed by 'assassinates' resists reductive analysis, so that one cannot independently specify just when killing counts as assassinating But it would be fallacious to infer on these grounds that assassinating is not killing Similarly, if the concept expressed by 'knows how' resists reductive analysis, then one cannot independently specify just when knowing-that counts as knowing-how But it would be nevertheless fallacious to infer that knowing-how is not knowing-that Again, the arithmetic concept expressed by 'is less than' cannot be defined in terms of the concept expressed by 'is distinct from' Nevertheless, being less than a number is a way of being distinct from it

---

[35] "The Problem of the Essential Indexical," *Noûs*, xiii (1979) 3-21

[36] That ways would have to be linguistically describable in a manner in which the one to whom knowledge-how is ascribed would recognize is assumed in an argument against a variant of our view in Ziff, p 70

[37] Thanks to Ian Rumfitt for pressing this concern

Perhaps an analogy to another kind of knowledge-that will help John's knowledge that he is tired is knowledge-that If one accepts modes of presentation, then such knowledge involves thinking of oneself under a first-personal mode of presentation There may very well be no way to analyze away John's first-personal mode of presentation of himself in third-personal terms But John's knowledge that he is tired is nevertheless genuine knowledge-that To say that first-person knowledge is a distinctive subclass of knowledge-that is not to deny that it is genuine knowledge-that The very same point applies to knowing-how Perhaps one cannot give a reductive analysis of thinking of a way under a practical mode of presentation But this does not undermine the thesis that knowing-how is knowledge-that [38]

But another version of this concern remains for our project According to us, knowing-how is a distinctive kind of knowing-that If the special subclass of knowing-that which we call 'knowing-how' is too dissimilar from other kinds of knowing-that, then one might suspect that we have just recreated the traditional distinction between knowing-how and knowing-that, but in other terms So it must be that, on our analysis, knowing-how possesses the characteristic features of other kinds of knowing-that This leads us to a third worry with our proposal

---

[38] Some philosophers reject modes of presentation altogether in the analysis of propositional attitudes, and so would not accept the Fregean arguments we have provided in the previous section for the existence of practical modes of presentation Practical modes of presentation are not essential to our analysis, in the sense that a philosopher who rejects modes of presentation simpliciter may easily accept the rest of our analysis of knowing-how For someone in this situation, this second worry of course does not arise at all

On the analysis we presented in the last section, knowing-how is analyzed in terms of knowing-that. In particular, knowing how to F is a matter of knowing that p, for a certain proposition p (as well as entertaining it under the right mode of presentation). So, knowing-how is straightforwardly analyzed in terms of knowing-that. But one might worry that significant disanalogies still remain between knowing-how and other kinds of knowing-that. One potential source of disanalogy involves Gettier cases.[39] We can imagine cases of justified true belief that fail to be knowledge-that, because they fail to satisfy some extra condition. It may appear difficult, however, to formulate examples that fall short of being knowledge-how for a similar reason. That is, one might think it is difficult to conceive of Gettier-cases for knowledge-how. But if knowledge-how is really a kind of knowledge-that, there should be such cases.

We doubt that every kind of knowledge-that is susceptible to Gettier cases. So it would not worry us if it were not possible to come up with a Gettier case for knowledge-how. Be that as it may, there are indeed Gettier cases for knowledge-how. Bob wants to learn how to fly in a flight simulator. He is instructed by Henry. Unknown to Bob, Henry is a malicious imposter who has inserted a randomizing device in the simulator's controls and intends to give all kinds of incorrect advice. Fortunately, by sheer chance the randomizing device causes exactly the same results in the simulator as would have occurred without it, and by incompetence Henry gives exactly the same advice as a proper instructor would have done. Bob passes the course with flying colors. He has still not flown a real plane. Bob has a justified true belief about how to fly. But there is a good sense in which he does not *know* how to fly.

Here is a fourth worry with our proposal. There are numerous differences between constructions such as

(29) Hannah$_i$ knows [how PRO$_i$ to ride a bicycle]

and constructions such as

(34) (a) Hannah knows how Bill rides a bicycle.
  (b) Hannah knows how Hannah rides a bicycle.
  (c) Hannah knows how one should ride a bicycle.

One might think that these differences provide evidence for the thesis that, while (34a-c) are ascriptions of knowledge-that, constructions such as (29) are not.[40] But the numerous differences between

---

[39] Thanks to Alex Byrne and Dean Pettit for raising this concern.

[40] For example, Ziff, who emphasizes the distinctions between (29) and (34c)—see p. 70.

(29), on the one hand, and (34a-c), on the other, all can be explained without threatening in the least the thesis that knowledge-how is a species of knowledge-that. Uses of (29) are associated with practical modes of presentation of ways of engaging in actions. In contrast, (34a-c) are not conventionally linked to practical modes of presentations of ways of engaging in actions. (29) involves a modal element missing in the sentences in (34a-b), and different from the modal element in (34c). (34a-b) entail, respectively, that Bill rides a bicycle, and that Hannah rides a bicycle, (29) carries no such commitment.

Other relevant differences between (29) and (34a-c) are traceable to facts about the interpretation of 'PRO' in untensed clauses. For example, uses of 'PRO' where they are controlled by the subject in the main clause invariably give rise to "de se" readings, that is, readings involving a first-person mode of presentation.[41] So, (29) ascribes self-knowledge, whereas (34a) and (34c) do not, and (34b) may or may not, depending upon context.

Therefore, the many differences between (29), on the one hand, and (34a-c), on the other, in no way undermine the thesis that knowledge-how is a species of knowledge-that.

Here is a fifth worry with our proposal. In many languages two meanings of the English word 'know' are lexically distinguished. The two meanings are those found in

(35) Hannah knows that penguins waddle.
(36) Hannah knows Bill.

For example, in German, the verb for the use of 'know' found in (35) is '*wissen*' and the verb for the use of 'know' found in (36) is '*kennen*'. In French, the verb for the use of 'know' in (35) is 'savoir', and the verb for the use of 'know' in (36) is '*connaître*'. This suggests that, in English, the word 'know' has two distinct lexical meanings, corresponding to the meanings that would be expressed in German by '*wissen*' and '*kennen*'.

---

[41] For example, suppose Hannah sees a picture of a woman in the newspaper who appears to be buying a lottery ticket, and furthermore appears greatly burdened by poverty. On this basis, she forms the desire that the woman in the newspaper win the lottery. Hannah herself is quite happy with her middle-class salary, and rightly suspects that the accumulation of more money would make her into an unpleasant person. Unbeknownst to Hannah, however, the woman in the newspaper is Hannah herself, who was buying a pack of cigarettes, rather than buying a lottery ticket. With respect to this context,

(a) Hannah wants to win the lottery.

has only a false interpretation. Two facts explain the lack of a true interpretation. First, the structure of (a) is

(b) Hannah$_i$ wants [PRO$_i$ to win the lottery].

Second, 'PRO', when controlled by the subject in the main clause, permits only de se readings.

There is internal evidence for an ambiguity in the English word 'know' between its use in (35) and its use in (36) The sentence

(37) John went to the bank, and Bill did too

cannot be used to express the proposition that John went to the riverbank, and Bill to the financial institution Such examples strongly suggest that meanings are preserved in ellipsis But consider

(38) ?* Hannah knows that penguins waddle, and Bill, Ted

(38) is clearly ungrammatical One explanation is that the elided expression has the meaning of the word 'know' in uses such as (35), but such a use requires a propositional complement, which it lacks in the second conjunct of (38) Examples such as (38) already suggest that the word 'know' is ambiguous in English between its uses in (35) and (36)

The worry with our proposal is that a similar point holds of the uses of 'know' in

(2) Hannah knows how to ride a bicycle
(3) Hannah knows that penguins waddle

Might evidence from English and other languages show that 'know' has different senses in (2) and (3)?

The worry is easily assuaged In the languages with which we are familiar, the uses of 'know' in (2) and (3) are translated by the same word This strongly militates against an ambiguity between the uses of 'know' in (2) and (3) [42] Moreover, ellipsis tests within English suggest that there is no such ambiguity in the English word 'know' Consider

(39) (a) Hannah knows that penguins waddle, and Bill, how to imitate them
(b) Bill knows how to ride a bicycle, and Hannah, that doing so is dangerous

(39a-b) are perfectly well formed If 'know' had different meanings in (2) and (3), however, sentences such as (39a-b) would be ill formed The crosslinguistic data, together with this evidence, demonstrate the nonexistence of the alleged ambiguity

---

[42] In certain languages, such as German, '*wissen wie*', the translation of 'to know how', must always take a tensed clause as a complement As a consequence, there is no direct German translation of constructions such as (29), though there are direct translations of constructions such as (34a-c) This distinction between English and German in no way indicates some deep conceptual difference between the English 'know how' and the German '*wissen wie*' It merely reflects the brute syntactic fact that German embedded questions cannot occur in untensed clauses, no matter what the question-embedding verb may be

Here is a sixth worry with our proposal Intuitively, ascriptions of knowledge-that are opaque That is, intuitively, (40) does not entail (41)

(40) Hannah knows that Hesperus is Hesperus
(41) Hannah knows that Hesperus is Phosphorus

One might worry, however, that ascriptions of knowledge-how are transparent rather than opaque That is, one might worry that inferences like that from (40) to (41) seem valid for cases of knowledge-how

But ascriptions of knowledge-how do not seem transparent For example, (43) does not seem to follow from (42)

(42) Hannah knows how to locate Hesperus
(43) Hannah knows how to locate Phosphorus [43]

Similarly, to modify an example of Carr,[44] suppose that Hannah, a famous dancer, knows how to dance a performance which she has dubbed 'Headbanger' Unbeknownst to her, her performance of Headbanger mirrors an accurate semaphore performance of "Gray's Elegy" which has become known as "Harvey " In this situation, (45) does not seem to follow from (44)

(44) Hannah knows how to perform Headbanger
(45) Hannah knows how to perform Harvey

Of course, as with ascriptions of knowledge-that, there is much room for pragmatic latitude in our judgments about ascriptions of knowledge-how But the issues seem no different for ascriptions of knowledge-how and ascriptions of knowledge-that [45]

Here is a seventh objection to our proposal On our account, knowledge-how is propositional knowledge But in certain situations, we smoothly ascribe knowledge-how to animals For instance, if Pip is a dog, someone might easily say

(46) Pip knows how to catch a Frisbee

---

[43] Peter Ludlow has suggested to us that the intensionality in (42) and (43) is due to the fact that 'locate' is an intensional transitive verb His suggested replacement for (42) is 'Hannah knows how to blow up Hesperus', which does not entail that Hannah knows how to blow up Phosphorus (perhaps she can only aim her rockets during a certain time of day)

[44] "The Logic of Knowing How and Ability," p 407

[45] It may be that embedded questions seem less opaque than clausal complements not headed by question words For example, 'John knows where Cologne is' seems less opaque than 'John knows that Cologne is in Germany' But this difference does not bear on the propositional nature of the construction Thanks to King for discussion here

One might think that nonhuman animals are not sufficiently conceptually sophisticated enough to possess propositional knowledge

But this objection is a nonstarter For in similar scenarios, we just as smoothly ascribe propositional knowledge to nonhuman animals, as in

(47) (a) Pip knows that when visitors come, he has to go into the kitchen
(b) Pip knows that Alva will give him a treat after dinner

So, smooth ascriptions of knowledge-how to nonhuman animals are simply no objection to our account Everyone requires some account of uses of sentences such as (47a-b) Whatever account is provided will work equally well for uses of sentences such as (46)

Furthermore, the possibility of ascribing knowledge-how to *F* goes with the possibility of ascribing false beliefs about how to *F*, which are clearly propositional For example, we might say

(48) The elephants know how to cross the river

They go to the only ford and walk across But now the river is dredged, unbeknownst to the elephants, and the ford disappears (48) becomes false, what is true is

(49) The elephants have a false belief about how to cross the river

Here is a eighth and final objection to our proposal We can formulate it using the following quote from Bechtel and Abrahamsen, which, together with the appeal to the alleged linguistic distinction discussed in the last section, is their sole argument for a distinction between knowledge-that and knowledge-how

> A person who *knows that* Sacramento is the capital of California will be able to retrieve from memory the proposition Sacramento is the capital of California, or to retrieve other propositions from which this one can be deduced But the same does not seem to hold for *knowing how* to ride a bicycle In this example, what is required is to have a certain ability to control one's perceptual-motor system (*op cit*, p 152)

Now, we have already seen that it is incorrect to analyze knowledge-how in terms of abilities, so we may discard Bechtel and Abrahamsen's positive account of knowledge-how Their quote nonetheless suggests an argument against our account According to this argument, knowledge-that requires a capacity to retrieve a proposition from memory, whereas knowledge-how does not

We doubt that knowledge-that requires a capacity to retrieve a proposition from memory We are not occurrently aware of many propositions that we know Examples include tedious geographical truths buried in one's memory External factors may prevent one from retrieving propositions from memory that entail them One may be preoccupied driving a car, or by the state of one's marriage None of this undermines one's knowledge of tedious propositions So, in order for the premise to have some initial plausibility, the possibility of retrieval from memory must be limited to certain favored circumstances

But if the premise is that, in certain favored circumstances, one can retrieve one's knowledge-that from memory, then there is no argument against the thesis that knowledge-how is just knowledge-that If the premise is plausible for knowledge-that, then it is just as plausible for knowledge-how For Hannah to know how to ride a bicycle, in certain favored circumstances, she must be able to retrieve some propositions expressed by sentences of the form '*w* is a way for Hannah (herself) to ride a bicycle' The favored circumstances may include sitting on a bicycle, and Hannah can retrieve the proposition without being able to express it in nonindexical words If one accepts the premise for knowledge-that, then one should accept it for knowledge-how

So much for objections to our account, now we turn to some of its benefits Its most obvious benefit is that it is the account entailed by current theories about the syntax and semantics of the relevant constructions Rejecting it would involve revising many well-entrenched beliefs about them in linguistics This move would be legitimate if the account could be shown to face serious difficulties But we have been unable to uncover such difficulties

A second benefit of the account is that it explains features of ascriptions of knowledge-how that other accounts leave unexplained For example, it has often been noted that people do not know how to do every sort of action As Ziff (*op cit*, p 71) points out

(50) Human babies know how to suck

seems false Although human babies do suck, presumably they do not know how to suck Similarly, it seems false to say that

(51) Human babies know how to cry for hours

Our account explains these facts On our account, knowledge-how is a species of knowledge-that And we are often reluctant to ascribe propositional knowledge to babies[46]

---

[46] According to Charlotte Katzoff—"Knowing How," *Southern Journal of Philosophy*, XXII (1984) 61-69—such attributions of knowledge-how are infelicitous, because false According to Katzoff, sucking is a *basic action* in Alvin Goldman's sense Furthermore, on her view, basic actions are not manifestations of knowledge-how

Of course, we are not always reluctant to ascribe propositional knowledge to babies For example, suppose Isobel is an infant One might smoothly say

(52) Isobel knows that when Richard comes home, she will be tossed in the air

Similarly, in certain situations, we may smoothly ascribe the problematic instances of knowledge-how to babies For example, suppose that there is some question about whether the infant Paul is handicapped The doctor may note with pleasure that

(53) Paul knows how to suck

If knowledge-how is a species of knowledge-that, then these parallels are no accident

Our account of knowledge-how both follows from basic facts about the syntax and semantics of ascriptions of knowledge-how and explains some distinctive features of ascriptions of knowledge-how that other accounts leave unexplained Furthermore, it is consistent with the theoretical significance of the intuitions that have motivated philosophers to reject our thesis In the next section, we turn very briefly to the consequences of our arguments for uses that philosophers have made of knowledge-how outside of epistemology

IV CONCLUSION

Outside epistemology proper, philosophers have made many uses of the thesis that knowledge-how is not a species of knowledge-that We conclude by considering two, one in the philosophy of mind and one in the philosophy of language In both cases, reliance on the alleged distinction between knowledge-how and knowledge-that is fatal to the thesis advanced Our brief discussion of these representative cases will underscore the dangers of invoking Ryle's distinction [47]

We disagree with Katzoff on this issue We think that basic actions can be manifestations of knowledge-how For example, one can sensibly ask a brain-damaged patient whether she still knows how to raise her arm Knowledge of the relevant way of raising one's arm is then demonstrated simply by raising one's arm

[47] There are, of course, numerous other uses of the alleged distinction between knowledge-how and knowledge-that One example is Adrian Moore, who uses the alleged distinction to argue that there is ineffable knowledge (for criticism, see Williamson, Review of Moore's *Points of View*, *Philosophical Books*, XL (1999) 43-45) Another example is John Searle's discussion of "The Background" in his *Intentionality* (New York Cambridge, 1983) One quite recent example is Stephen Yalowitz—"A Dispositional Account of Self-Knowledge," *Philosophy and Phenomenological Research*, LXI, 2 (2000) 249-78—who makes essential appeal to a nonpropositional notion of knowledge-how in responding to Crispin Wright's arguments for a noncognitivist account of semantic self-knowledge (cf pp 270-71) There are more

One particularly well-known use of Ryle's account of the relation between knowledge-how and knowledge-that occurs in Lewis's reply to Frank Jackson's famous “knowledge argument” Jackson imagines Mary, a brilliant scientist locked up in a room her entire life, who only sees television images of the world in black and white Mary develops a full mastery of all natural sciences, including physics and neurobiology In particular, Mary has a full knowledge of the neurophysiology underlying color vision and the physics of color But when Mary leaves her black and white room, and sees red for the first time, intuitively, she acquires new knowledge If so, according to Jackson, not all knowledge-that is knowledge of propositions about the physical world

According to Lewis, the correct account of Jackson's knowledge argument is that Mary does not gain new knowledge-that when she leaves her black and white room, but only knowledge-how In particular, she gains knowledge how to recognize, remember, and imagine experiences of red Our discussion shows, however, that Lewis's account is incorrect Knowing how to imagine red and knowing how to recognize red are both examples of knowledge-that For example, *x*'s knowing how to imagine red amounts to knowing a proposition of the form '*w* is a way for *x* to imagine red', entertained under a guise involving a practical mode of presentation of a way

There is perhaps a fallback position available to Lewis According to it, all we have shown is that Lewis has misappropriated the expression 'knowledge-how' If so, then Lewis can simply give up this locution, and recast his account purely in terms of abilities On this account, what Mary gains is not knowledge-how, but rather simply new abilities Indeed, this “ability” analysis of the knowledge argument is the now standard response to the knowledge argument It occurs not only in Lewis, but also in quite influential work by others [48]

There are two ways to develop this fallback position According to the first, there is no knowing how to imagine an experience of red There is just being able to imagine an experience of red If so, then Mary does not gain any new knowledge-how, but only a new ability But this position is not a plausible response to Jackson's argument against physicalism For the ability to imagine an experience of red is clearly an ability to perform an intentional action And we do find it very plausible that intentional actions are employments of knowledge-how. Indeed, as discussed in section I, the thesis that intentional actions are employments of knowledge-how is precisely what accounts for the initial plausibility of Ryle's original argument against the claim that knowledge-how is a species of knowledge-that. But if intentional actions are in fact employments of knowledge-how, then Mary's acquisition of an ability to imagine an experience of red brings with it knowledge how to imagine red, and with it, a bit of propositional knowledge that she previously lacked.

---

[48] In particular, see Janet Levin, “Could Love Be Like a Heatwave? Physicalism and the Subjective Character of Experience,” *Philosophical Studies*, XLIX (1986) 245-61, and Laurence Nemirow, “Physicalism and the Cognitive Role of Acquaintance,” in Lycan, ed , pp 490-99

Indeed, independently of any general thesis about intentional action, it is plausible that the ability to imagine an experience of red entails knowing how to imagine an experience of red. For example, intuitively, the best explanation of why Mary lacks the ability to imagine an experience of red in her black and white room is that she does not know how to do so.

The second way to develop this fallback position is to grant to Mary, in her black and white room, knowledge how to imagine an experience of red. What Mary then gains upon emerging from her black and white room is the ability to employ this knowledge. Our problem with this response is straightforward. It seems absurd to countenance the truth of

(54) Mary knows how to imagine an experience of red

with respect to the situation in which Mary is in her black and white room. If she knows how to imagine an experience of red, why is she unable to imagine such an experience? Evidence for the robustness of the intuition that Mary does not know how to imagine an experience of red is the fact that, throughout the literature, the falsity of (54) with respect to the envisaged situation is assumed.

Therefore, the ability account of Jackson's knowledge argument fails to show that Mary does not acquire propositional knowledge that she did not previously possess upon leaving her black and white room. Indeed, assuming that Mary did not possess the requisite knowledge-how already, the ability account in fact *entails* that Mary acquires propositional knowledge upon leaving her black and white room.

Another use of Ryle's distinction between knowledge-how and knowledge-that occurs in Devitt's discussion of linguistic competence (*op. cit.*), where one of his central purposes is to undermine a doctrine he calls "Cartesianism." This doctrine is characterized in several nonequivalent ways. But one way in which it is characterized, to which Devitt repeatedly returns, is that linguistic competence with a term entails possession of propositional knowledge about its meaning (*ibid.*, pp. 26-27, 52, note 4, 173). Devitt attempts to undermine this view by appeal to Ryle's distinction. According to Devitt, someone who knows the meaning of a term knows how to use that term with a certain meaning. But this knowledge-how does not thereby grant the user of the term any propositional knowledge. It follows, according to Devitt, that "There is no good reason to suppose that a person who is competent with a sentence—who has the ability to use it with a certain meaning—must thereby have any propositional knowledge about what constitutes its meaning" (*ibid.*, p. 173).

Devitt is never very explicit about what sort of knowledge-how is to be identified with competence with a term. But, as the last quotation suggests, competence with a term *t* at least involves knowing how to use *t* with a certain meaning, presumably the meaning it actually has. But if competence with a term *t* involves knowing how to use *t* with the meaning it actually has, then linguistic competence with a term does, on Devitt's own characterization, yield propositional knowledge about the meaning of that term. For, given a term *t* which has a certain meaning *m*, *x*'s knowing how to use *t* with the meaning *m* amounts, for some contextually relevant way *w*, to *x*'s knowing that *w* is a way for *x* to use *t* with the meaning *m*. So, Devitt's own characterization of competence commits him to the doctrine he labels "Cartesianism."

We began this discussion by rejecting the original arguments that motivated the alleged distinction between knowing-how and knowing-that. We then presented our own account of knowledge-how, according to which it is a species of knowledge-that, and justified it by appeal to well-entrenched doctrines of linguistic theory. Following this, we showed how our account smoothly explains all of the different phenomena that have led philosophers to embrace a false dichotomy between knowing-how and knowing-that.

All knowing-how is knowing-that. Neglect of this fact impoverishes our understanding of human action, by obscuring the way in which it is informed by intelligence.

JASON STANLEY

University of Michigan

TIMOTHY WILLIAMSON

Oxford University

THE JOURNAL OF PHILOSOPHY

VOLUME XCVIII, NO. 9, SEPTEMBER 2001

THE GROUNDLESS NORMATIVITY OF INSTRUMENTAL RATIONALITY*

In *The Little Prince*, Antoine de Saint-Exupery[1] has us imagine a monarch who, wanting above all to be obeyed, tailors his commands to the anticipated behavior of his subject. When the little prince yawns, the King commands him not to do so again. The little prince objects—pleading that he cannot help yawning—so the king commands the little prince *to* yawn. Frightened, the little prince is unable to yawn. The king struggles for a moment and then commands the little prince to "sometimes yawn and sometimes not." As Saint-Exupery says:

> For what the king fundamentally insisted upon was that his authority should be respected. He tolerated no disobedience. He was an absolute monarch. But, because he was a very good man, he made his orders reasonable (*ibid.*, p. 37).

When the king boasts that he is the absolute ruler over all the universe and that even the stars and the planets obey him, the little prince asks the king to command a sunset. To this seemingly modest request of one with such power, the king replies:

> "If I ordered a general to fly from one flower to another like a butterfly, or to write a tragic drama, or to change himself into a sea bird, and if the general did not carry out the order that he had received, which one of us would be in the wrong?" the king demanded. "The general, or myself?"

---

* A version of this paper was read at a colloquium at Bowling Green State University. I am grateful to the audience for helpful suggestions. I am especially grateful to the participants in the Ohio Reading Group in Ethics—in particular, Justin D'Arms, David Copp, David Sobel, and Sigrún Svavarsdóttir—for careful reading and extremely useful comments.

[1] New York: Harcourt, Brace and World, 1943.

0022-362X/01/9809/445–68

> "You," said the little prince firmly
>
> "Exactly One must require from each one the duty which each one can perform," the king went on "Accepted authority rests first of all on reason If you ordered your people to go and throw themselves into the sea, they would rise up in revolution I have the right to require obedience because my orders are reasonable "
>
> "Then my sunset?" the little prince reminded him for he never forgot a question once he had asked it
>
> "You shall have your sunset I shall command it But, according to my science of government, I shall wait until conditions are favorable "
>
> "When will that be?" inquired the little prince
>
> "Hum! Hum!" replied the king, and before saying anything else he consulted a bulky almanac "Hum! Hum! That will be about—about—that will be this evening about twenty minutes to eight And you will see how well I am obeyed" (*ibid*, p 38)

I INTRODUCTION

What are we to make of the king and his commands? Were the king to refrain from giving commands only when the subject is incapable of complying, we would think this, in general, admirable and reasonable To order what one knows to be beyond the capacity of an individual may sometimes have a point, but the point can never be the standard one of trying to bring about compliance Indeed, were the king also to refrain from commanding those actions which impose too great a cost on his subjects—ordering them into the sea, for example—we might again think this a laudably restrained authority (at least provided nothing of overwhelming importance required such a sacrifice from his subjects) But, of course, the limitations the king places on his commands are, in fact, quite a bit more substantial than these What the king refuses to do is to command (or at least persist in commanding) any action that *will* not be performed As a result, while he is never disobeyed, neither does he (at least in the examples we see) exercise any authority at all

The commands of the Saint-Exuperian king, while in the form of prescriptions, are too "reasonable" to be regulative Rather than require that actions be adjusted to his commands, the king carefully tailors his commands to the actions that would be performed anyway His commands are not normative in any interesting way "Unreasonability" in the Saint-Exuperian sense is a requirement on a normative theory in order to avoid what might be called *normative triviality*

To some, it has appeared that Humean instrumentalism—perhaps the most influential theory of practical rationality—is as bumbling and ineffective in issuing practical norms as Saint-Exupery's king.[2] Humean instrumentalism grounds an agent's reasons for acting on the agent's desires, values, or some other subjective, contingent, conative states of the agent. The concern is that by doing this, the theory renders impossible the sort of normativity that a theory of practical rationality must have. Humean instrumentalism fails a criterion of adequacy for a theory of practical reasons—that of being "unreasonable" (in the Saint-Exuperian sense).

I argue, on the contrary, that the Humean instrumentalist can be as normatively demanding as it is evident that our judgments about practical reasons require. The critics have a complaint in this regard, but it is a question-begging complaint; it appeals to intuitions that the Humean does not, and ought not, share.

But the allegation of normative triviality is, I believe, better understood as a symptom of a different concern about Humean instrumentalism. What provokes the charge of normative triviality is the concern that Humean instrumentalism has its feet firmly planted in the air—that it offers reasons that are, in the end, ungrounded. Critics have thought that, at best, the Humean could offer an account of the transmission of practical reasons, never of their origin. So long as the Humean tells us that practical reasons are transmitted from ends to means, she is an instrumentalist. For reasons to be transmitted, however, they must exist in the first place. There must be reasons for the ends—at least those which are "ultimate" ends. And, to justify the claim that there are such reasons, the Humean must "go beyond" instrumentalism—she must commit herself to the existence of reasons that are not conditioned on the subjective, contingent, conative states of the agent.

This allegation, I argue, is based on a fundamental misunderstanding of the nature and purpose of Humean instrumentalism. On a proper construal, the theory adroitly side-steps these attacks. We begin, though, with the charge of normative triviality.

II. NORMATIVE TRIVIALITY

It is not completely clear how, precisely, to formulate the requirement that a theory be normatively demanding—that it avoid normative triviality. The story of the little prince might suggest that the problem arises when apparent demands are always satisfied, these "commands" never require actions that are, in fact, not done But obviously, it cannot be a requirement on a normative theory that its prescriptions actually be violated This would be to make a mistake opposite to that made by Saint-Exupery's king

---

[2] What I am here referring to as *Humean instrumentalism* I have elsewhere called, and shall later in this essay call, *neo-Humean instrumentalism*. The qualification is desirable to contrast it with the views that might plausibly be attributed to Hume himself. For purposes of introducing the issues here, I am setting aside my preferred terminology in order to speak in terms that are more common. In these commonly used terms, Hume seems not to be a Humean.

Suppose we require that a (nontrivial) prescription be one that the agent is capable of violating This seems too restrictive Surely, I can be guided by a norm, even if I do not have the capacity to act contrary to it Cases of overdetermination suggest this conclusion Imagine that I am one of Stanley Milgram's[3] real subjects in his famous experiments on obedience to authority Suppose I am told to increase the electrical shock I apply to the supposed subjects but, guided by a norm not to harm innocent people, I refuse Now, no matter what I had done—press the button or not—I would not have been able to harm the apparent subject It was not within my capability to do so It seems to me, however, that I was guided by the norm in question in this case, even though I did not have the capability to violate the norm

We could say that the *minimal demandingness* necessary to avoid normative triviality requires that it must be possible to violate a norm [4] A putative norm is normatively trivial when it is not possible for an agent to violate the norm But this runs into the overdetermination problem, too It seems to me that, if I do not know that some apparent act description in fact refers to no possible action, I can be "guided by" a norm that condemns such "actions " For example, after listening to a logical proof given by a colloquium speaker, I might—falsely believing the demonstration fallacious but thinking it bad form to refute a guest in public—be guided by a norm not to demonstrate publicly the logical fallacies committed by colloquium speakers, even though it is logically impossible for me to violate this norm in this case

I do not want to track this problem very far because I do not think precision in the formulation of the nontriviality requirement is necessary for my purposes It seems to me, though, that the problem with Saint-Exupery's king's commands is a "direction of fit" problem As I suggested before, the king tailors his commands, as much as he is able, to his expectations about the future actions of his subjects—indeed, to his expectations about the actions that his subjects would perform independently of his commands The explanation of why the king is never disobeyed runs from his expectations of their actions to the content of his commands When norms are authoritative, they are not created to match independent expectations about the actions they "govern "[5]

---

[3] *Obedience to Authority* (New York Harper and Row, 1974)

[4] This seems to be the sort of requirement Christine Korsgaard endorses, as discussed below

III NORMATIVE TRIVIALITY AND SUBJECTIVE EXPECTED UTILITY

Does the requirement that I have called (minimal) demandingness have any bite with respect to theories of practical rationality? It might Some have suggested that one standard instrumentalist theory, the *subjective expected utility* (SEU) theory of practical rationality, does not satisfy the requirement of demandingness The reasoning goes as follows

SEU holds that one acts rationally when one acts so as to maximize one's expected utility, where that notion is a measure of the probabilistically weighted utility of the various possible outcomes of the action On this account, the probabilities, as well as the utilities, are measures of some subjective state of the agent Probabilities are measures of the agent's subjective degree of belief and utilities are measures of her subjective preference But how do we determine an agent's preferences and degrees of belief?

For expository reasons, let me drop the second part of this question and focus on the measurement of preference One approach to measuring an agent's subjective preferences—one that is still popular in some quarters—is based on the *doctrine of revealed preference* This doctrine holds that "preference is revealed in choice " The slogan is misleading It could refer to a rather modest epistemic thesis that choice behavior is evidence—perhaps, in many cases, the best evidence we have—of an agent's preferences In fact, though, it is a metaphysical thesis preference just is choice behavior, or, more plausibly, dispositions toward choice behavior

In a simple dichotomous choice situation, an agent prefers A to B just in case she takes (or would take) A rather than B when offered an exclusive choice Making certain diachronic consistency assumptions, behavior (or dispositions to behave) over a suitably large number of pairwise choices, allows the attribution of measures of utility that are unique up to a positive linear transformation And such measures are taken to be strong enough for doing the work of decision theory

---

[5] While I think the issue of triviality is ultimately to be cashed out in terms of "direction of fit," I shall often speak as if the issue is simply whether the norm might require some action that would not be done independently of the norm This is for ease of presentation only

Now, if we use utilities, so constructed, as a basis for a normative SEU theory, a problem arises. Amartya Sen[6] puts it this way:

> If you are observed to choose *x* rejecting *y*, you are declared to have "revealed" a preference for *x* over *y*. Your personal utility is then defined as simply a numerical representation of this "preference," assigning a higher utility to a "preferred" alternative. With this set of definitions you can hardly escape maximizing your own utility, except through inconsistency (*ibid.*, p. 322).

But the matter is worse than Sen recognizes. Behavior is a part of nature and nature does not permit contradictions. Unless we make substantive assumptions about an agent's preference structure—assumptions which cannot be warranted on any purely formal grounds or on strictly behavioral grounds—we cannot convict an agent of inconsistency.

Clearly, what Sen has in mind when he excepts inconsistency is the case of the person with so-called incoherent preferences—for example, a person with a cyclic preference pattern. Such a person prefers *A* to *B*, *B* to *C*, and *C* to *A*. This sort of cyclic preference pattern is usually alleged to be irrational on the grounds that an agent with such a preference pattern can be turned into a money pump—paying some small amount of a good, say $1.00, to make each of the following exchanges: *B* for *C*, *A* for *B*, and *C* for *A*. She winds up "where she began but three dollars poorer."

Such a conviction, though, cannot be maintained on the strict behaviorist grounds offered by the doctrine of revealed preference. A simple example will illustrate well enough. Imagine what I shall call an "All-You-Can-Enjoy Spa," consisting of a Jacuzzi, a sauna, and a swimming pool. One may, let us suppose, enter any room for the price of a dollar. Now imagine that, tired after a hard day's work, Arnold pays a dollar to sit in the Jacuzzi for twenty minutes. Then, for a dollar, he moves to the sauna for a few minutes before paying yet another dollar to enter the swimming pool for a quick workout. Finally, tired from his workout, he pays a dollar to enter the Jacuzzi again.

We do not think that Arnold is "back were he started but three dollars poorer"—at least not in any sense that would lead us to accuse him of being a "money pump." We can perfectly well make sense of the preferences that "rationalize" this behavior. We do not think he has done anything irrational or expressed inconsistent or incoherent preferences. But, behaviorally, there is nothing to distinguish him from a classical money pump. Our only reason for condemning the classical money pump as displaying practical irrationality is that we balk at the implications of revealed-preference theory. In the stories told about money pumps, we refuse to think that there is an intelligible, coherent set of preferences that rationalize the behavior. But think it we must if we are inferring preferences from behavior as the doctrine commands us. So, perhaps the classical money pump desires to own each of the objects, *A*, *B*, and *C*, for some short period of time. Making the apparently cyclic trades, then, is no more irrational than is Arnold's moving through the various rooms of the spa.

---

[6] "Rational Fools," *Philosophy and Public Affairs*, vi (1977): 317-44.

If we accept the doctrine of revealed preference, taking behavior not merely to be evidence of an agent's preferences but to be constitutive of the agent's preferences, the conclusion about the inevitability of maximizing expected utility is stronger than Sen realizes: with this set of definitions, you cannot escape maximizing your own utility, period. So, if we develop an instrumentalist normative theory of practical rationality along the lines of a SEU theory that understands utility as a measure of the agent's subjective preferences and these preferences to be constituted by the agent's behavior, our theory will be normatively trivial. It is "reasonable" in Saint-Exupery's sense, and so, unreasonable as a normative theory.

But it is quite clear where lies the weak point in this cluster of claims—and it is not with the instrumental theory of practical rationality. Obviously, if we are to construct a normative theory of practical rationality along the lines of a SEU theory and we think that our theory must not be normatively trivial, then we cannot understand utility as being constituted by actual behavior. So much the worse for the doctrine of revealed preference as a definition of "utility."

But this account of "utility" has little to recommend it anyway. Its starkly behaviorist credentials—perhaps attractive to an earlier generation of philosophers—do not hold out much attraction now even if our objective is a predictive or explanatory theory of intentional human behavior. Choice can never "reveal" preference in any situation but the specific, concrete situation in which it is observed. Behavior alone never tells us what are the relevant features of the situation. Because of this, we do not know, on behaviorist grounds alone, how to generalize from observed behavior in order to predict future behavior. Without background, nonbehaviorist assumptions, no future behavior is rendered inconsistent, or even less probable, based on observed past behavior. The revealed preference version of SEU theory is not predictive, unless we make assumptions about agents' preferences which are not behavioristically based. And in what significant sense can it be explanatory to account for behavior in terms of utilities that are just summaries of behavior?

The doctrine of revealed preference is problematic all by itself It is not surprising that, when conjoined to an instrumentalist, preference-based theory of practical rationality, it results in a problematic theory Indeed, it would be quite remarkable if it did not To paraphrase J S Mill[7] there is no difficulty in proving any standard of rationality whatever to work ill if we suppose crude behaviorism to be conjoined to it It seems obvious that one who seeks an adequate instrumentalist, preference-based theory of practical rationality should look to some other account of "preference " But, lately there has been a growing concern that this is not possible More precisely, the thought that some have promoted is that a satisfactory account of preference will be inconsistent with an instrumentalist, preference-based theory of practical rationality

Christine Korsgaard[8] is one of the prime movers of this criticism According to Korsgaard, David Hume, like the revealed-preference theorist, is committed to the view that people cannot violate the instrumental principle Indeed, I should have said that, on Korsgaard's reading, Hume, like *other* revealed-preference theorists, is committed to the view that people cannot violate the instrumental principle, for Korsgaard interprets Hume as holding what amounts to a revealed-preference theory Let us grant the interpretive point[9]

---

[7] "Utilitarianism," in *Selected Writings of John Stuart Mill*, Maurice Cowling, ed (New York Mentor, 1968), for the original quote which concerns the need for usable decision procedures in moral practice, see p 265

[8] See, especially, "The Normativity of Instrumental Reason," in Garrett Cullity and Berys Gaut, eds , *Ethics and Practical Reason* (New York Oxford, 1997), pp 215-54 Warren Quinn is another philosopher who develops and promotes this attack on neo-Humean instrumentalism—see *Morality and Action* (New York Cambridge, 1993), pp 210-55

[9] I am dubious, though As we shall see momentarily, Korsgaard thinks the problems arise for Hume because, "Hume identifies a person's *end* with what he *wants most*, and the criterion of what a person wants most appears to be what he actually *does*" (*op cit*, p 230) In a very famous passage, however, Hume seems to recognize the possibility of failing to do what you want Here is the familiar passage

> '[T]is only in two senses, that any affection can be call'd unreasonable First, When a passion, such as hope or fear, grief or joy, despair or security, is founded on the supposition of the existence of objects, which really do not exist Secondly, When in exerting any passion in action, we chuse means insufficient for the design'd end, and deceive ourselves in our judgment of causes and effects—*A Treatise of Human Nature* (New York Oxford, 1968), p 416

Now it is, I think, quite inconsistent for someone simultaneously to believe that we may, in action, sometimes choose means insufficient for our designed ends and to believe, as Korsgaard suggests Hume does, that a person's ends are revealed in

What are its implications? Korsgaard believes that a principle it is impossible to violate cannot be normative, one cannot be guided by such a principle. As was argued above, this is not a satisfactory account of normative triviality. But assuming revisions along the lines suggested there—revisions to which she would probably not object—I agree with Korsgaard on this point.

Korsgaard uses this criterion, which I have called *demandingness,* to pose a problem for neo-Humean instrumentalism. The problem has the form of a dilemma within a dilemma. The first dilemma is this: either the neo-Humean instrumentalist's theory is normatively trivial, or it draws a distinction between an agent's ends and what the agent, in fact, does. The second, nested, dilemma is a constructive one seeking to show that whichever of two paths the neo-Humean takes to draw this distinction, he must "go beyond" an instrumentalist theory. So, overall, we are pushed to the conclusion that either neo-Humeanism instrumentalism is normatively trivial or, contrary to advertising, it is committed to its own noninstrumental, unconditional rational requirement—its own "categorical imperative."

IV. INSTRUMENTALISM AND NORMATIVE DEMANDS

Can neo-Humean instrumentalism be demanding? In discussing the potential *un*demandingness of the SEU theory, we located the source of the problem in the doctrine of revealed preference—a doctrine sometimes phrased so as to appear merely epistemic, but, in fact, metaphysical to the core. Not surprisingly, in her discussion of Hume's theory, Korsgaard does the same:

> The problem is coming from the fact that Hume identifies a person's *end* with what he *wants most,* and the criterion of what a person wants most appears to be what he actually *does*. The person's ends are taken to be revealed in his conduct. So, the problem would be solved if we could make a distinction between a person's ends and what he actually pursues (*op. cit.*, p. 230).

Can we make the necessary distinction? It would seem obvious that we can, and not merely ad hoc so as to satisfy the criterion of demandingness.

First, what may be a terminological complaint. Korsgaard holds that the distinction we need to make is between a person's ends and "what he actually *pursues*" (emphasis added). But the doctrine of revealed preference says nothing about what a person pursues (in any

---

his conduct. While it is quite possible that Hume holds inconsistent beliefs about this matter, charity bids us to come to this conclusion only reluctantly.

natural sense of 'pursues') Rather, the doctrine of revealed preference holds that preference is revealed in conduct While we could deny a distinction between what a person pursues and what she does, I think we should not When talking with her overbearing mother, Amelia repeatedly and predictably says things that provoke conflict This is part of her conduct But she may not intend to provoke conflict, she may not desire conflict, she may not pursue conflict Amelia does provoke conflict with her mother—and she does so as the result of her voluntary, intentional actions—but she does not pursue conflict with her mother [10]

If not everything that an agent does is something the agent pursues, then Korsgaard's argument does not show that, in order to satisfy the requirement of demandingness, the neo-Humean instrumentalist needs a distinction between an agent's ends and what she pursues Neo-Humean instrumentalism will be at least minimally demanding if it can draw a distinction between mere conduct and an agent's ends, even if it can draw no distinction between an agent's ends and what she pursues So, let us suppose that conflict with her mother is something that Amelia not only does not pursue, it is something she desires not occur Then, the neo-Humean instrumentalist can say that Amelia has a reason to refrain from making the inflammatory comments Neo-Humean advice is corrective, it is demanding This does not require a gap between Amelia's ends and what she pursues, it is enough that there be a gap between her ends and what she does

I doubt that this is demanding enough But, before moving to this point, let us take stock Provided the neo-Humean instrumentalist can distinguish between the actions required by the instrumentalist principle in service of an agent's ends and her actual behavior, he can satisfy the requirement of demandingness, as we have explicated it That is to say, it is simply not true, as Korsgaard says, that the sort of position we are now considering "exclude[s] the possibility that we could be *guided* by the instrumental principle" because "anything [we] do counts as following it" (*op cit*, p 229)

But is this demanding enough? Let us say that a principle is demanding *with respect to actions* if it could, under some conditions, prescribe actions that are not undertaken A neo-Humean instrumentalism that distinguishes between an agent's pursuits and what she actually does is demanding with respect to actions even if it cannot distinguish between an agent's ends and what she actually pursues. For all that I have said though, such a theory cannot be demanding *with respect to pursuits*. And, one might plausibly claim, an adequate theory of practical rationality must be demanding with respect to pursuits, as well.

---

[10] In some cases of this sort, we might correctly attribute an unconscious motivation to Amelia But such an attribution will be implausible in others All that is required for my claim to be true is that there are some cases in which one brings about a result by one's voluntary, intentional actions but does not pursue that result

I agree. An adequate theory of practical rationality should not only allow space between those actions we are rationally committed to by our ends and those actions we actually perform. It should also allow space between those actions we are rationally committed to by our ends and those actions we are rationally committed to by our pursuits. In other words, it should allow for the rational criticism of (at least some of) our pursuits.

Can a neo-Humean instrumentalist do this? Most assuredly. For many of our pursuits do not further our ends. Remember Amelia. Let us suppose now that what she is pursuing is harmony with her overbearing mother. But she pursues this in the belief that harmonious relations will eventually lead her mother to quit being so overbearing and, as a result, their relationship will develop into a healthier one. What she wants most is to have this sort of healthy relationship with her mother. In light of this, she sets about to control her tendency to say things that provoke conflict. She pursues harmony in the relationship. In fact, though, harmonious relations with her mother will never result in a healthy relationship. Instead, the only way to bring about the sort of change Amelia most wants in the relationship is for her to stand up for herself forcefully against her mother's intrusions in her life. Now, Amelia's pursuit of harmony does not further her ends. What is rationally advisable given her ends is for her to stand up for herself against her mother's intrusions.[11]

---

[11] If Amelia does not believe that her ends would be better promoted by giving up the pursuit of harmony and standing up for herself, then her failure to change her pursuits does not reflect a failure in her practical rationality. If she has no reason to believe that her ends would be better served by changing her pursuits and some reason to believe that they would not, then her failure to change her pursuits does not reflect on her rationality at all. Still, the advisable course of action in light of her ends is to change her pursuits. There is no reason that a neo-Humean cannot recognize this as a legitimate and interesting evaluation of actions from Amelia's evaluative point of view. This sort of evaluation, while it steps outside of Amelia's doxastic and epistemic situation, does not adopt a nonsubjective evaluative perspective, it evaluates actions, as any neo-Humean theory would, from the evaluative perspective of the agent. When Amelia chooses actions, she is *aiming* to choose actions that do promote her ends—not actions that she believes promote her ends or actions that it would be reasonable to believe would promote her ends given her epistemic situation. When Amelia solicits advice from others about how to act, she is seeking guidance about what *will* further her ends, not about what it is reasonable

Here, we have a gap between the actions normatively required by Amelia's ends and actions normatively required by at least some of what she pursues in the mistaken belief that it will secure her ends Amelia should have different pursuits

How can this distinction between ends and pursuits be drawn? One plausible thought is that a person's ends are those states of affairs the person values in themselves Pursuits include everything that the person sets about to make true Neo-Humean instrumentalism can be demanding with respect to pursuits by evaluating them in light of what the agent values in itself Perhaps even this is not demanding enough Some would hold that an adequate normative theory of practical rationality must be demanding with respect to an agent's ends

Before evaluating the plausibility of this requirement and the ability, if any, of the neo-Humean instrumentalist to meet it, it is worth noting that the ground has shifted We have, by now, moved from an almost completely uncontroversial criterion of adequacy for a normative theory of practical rationality—nontriviality—to one that is very nearly question begging Many neo-Humean instrumentalists would claim that it is precisely a tenet of their theory that an agent's ultimate ends—the things she values in themselves—are not subject to rational appraisal Insistence that a normative theory of practical rationality allow us to subject an agent's ultimate ends to rational criticism does not provide a criterion of adequacy for such a theory, it merely states a rejection of one prominent family of theories

While I do not find the requirement that a normative theory of practical rationality be demanding with respect to ends a very useful one in resolving disagreement about these theories, I do think the neo-Humean can meet even this requirement, at least as the requirement has been formulated There may, after all, be straightforward instrumental reasons for acquiring or extinguishing ultimate ends Amelia desires, for its own sake, to have a healthy relationship with her mother Achieving this end, however, would require Amelia to devote so much of her attention and effort to this task that she would not be able to achieve any of the other ends she has Now, by itself, this does not "speak against" the end of having a healthy relationship with her mother Such an end is simply one that Amelia cannot afford—or is unwilling—to pay for, given the costs involved Amelia

---

to believe will do so given her beliefs or given beliefs she should have in her epistemic situation I discuss these issues in "Irrational Desires," *Philosophical Studies*, LXII (1991) 23-44, especially p 25 finds, however, that so long as she intrinsically values (has as an end) a healthy relationship with her mother, she keeps devoting time and effort to bringing about this state of affairs, even though she knows it is hopeless. She "chuse[s] means insufficient for the design'd end." She would achieve those of her ends that she can achieve more fully if she did not intrinsically value a healthy relationship with her mother. In this case, instrumental rationality, as understood by the neo-Humean, recommends that Amelia extinguish (to the degree that she can without incurring too high a cost) her intrinsic desire to have a healthy relationship with her mother. Ultimate ends can be instrumentally irrational in the sense that one has reason to extinguish them. In this sense, the neo-Humean can subject even an agent's ultimate ends to rational scrutiny.[12] What the neo-Humean instrumentalist is unable to do is to find an agent's ultimate ends intrinsically rational or irrational.

Let us call a principle *instrumentally demanding with respect to ends* when it holds, based partly on instrumental considerations, that an agent's ultimate ends may be different from those she should, rationally, have. A theory that, based solely on intrinsic considerations, holds that an agent's ends may not be those she should rationally have is *intrinsically demanding with respect to ends*. Consider these properties as putative criteria of adequacy for a normative theory of rational choice. Now, my claim is that even neo-Humean instrumentalism satisfies the first proposed criterion. And, further, I claim that the second criterion is completely question-begging. It asks us to accept as a criterion of adequacy of a normative theory of practical rationality that a central tenet of the neo-Humean theory be denied. Hardly an Archimedean leverage point!

The Neo-Humean instrumentalist theory, unmodified, satisfies the noncontroversial requirements of demandingness provided it is not hobbled with a very implausible and outdated behavioristic account of an agent's ends. Indeed, the theory can satisfy far more stringent and controversial requirements of demandingness—including demandingness with respect to pursuits and instrumental demandingness with respect to ends. So, what is the problem?

V. KORSGAARD'S DILEMMA

Korsgaard believes that neo-Humean instrumentalists face two dilemmas, one nested in the other. The first is between normative triviality and distinguishing between an agent's ends and what she actually does What I have argued is that the neo-Humean instrumentalist (and, indeed, I think Hume, himself), rather obviously and easily, takes the second horn But, on the tip of this horn waits another dilemma Korsgaard presents this second dilemma by offering the neo-Humean instrumentalist two ways to draw the distinction necessary for avoiding normative triviality (But Humeans should beware of Kantians bearing gifts)

---

[12] See my "Irrational Desires" (pp. 23–26) for a fuller discussion of this issue with a somewhat different focus.

Here is how Korsgaard presents the second dilemma

> [T]he problem would be solved if we could make a distinction between a person's ends and what he actually pursues Two ways suggest themselves we could make a distinction between actual desire and rational desire, and say that a person's ends are not merely what he wants, but what he has reason to want Or, we could make a more psychological distinction between what a person thinks he wants or locally wants and what he "really wants" (*op cit*, p 230)

The first approach seems, rather obviously, to grant a role to rational appraisal that defenders of instrumentalist conceptions of practical rationality have rejected, it seems to entail that there can be intrinsic rational appraisal of an agent's ultimate aims But what I have offered above in defending the demandingness of neo-Humean instrumentalism can be seen as a version of Korsgaard's second approach to avoiding normative triviality And, I believe, the sort of theory I sketched is what most neo-Humeans have intended all along So, what sharp point does Korsgaard have for those who take the second horn of her second dilemma?

She says

> [T]he second option also takes us beyond instrumental rationality, although this may not be immediately obvious If we are going to appeal to "real" desires as a basis for making claims about whether people are acting rationally or not, we will have to argue that a person *ought* to pursue what he *really* wants rather than what he is in fact *going* to pursue That is, we will have to accord to these "real" desires some normative force It must be something like a requirement of reason that you should do what you "really want," even when you are tempted not to And then, again, we will have gone beyond instrumental rationality after all (*op cit*, p 230)

This charge is initially puzzling The alleged step beyond instrumental rationality—the assertion that "a person *ought* to pursue what he *really* wants"—is just a reassertion of the neo-Humean instrumental principle As David Lewis[13] says in another context, "what is already there cannot be *added*" (*ibid*, p 534), one does not take a step beyond neo-Humean instrumentalism by repeating the thesis of neo-Humean instrumentalism

But Korsgaard's complaint is that the neo-Humean instrumentalist, no less than her opponents, is committed to a substantive principle of rationality which generates reasons that are not contingent on the agent's aims and desires—one which is, in this sense, "categorical " The emphasis by neo-Humean instrumentalists on means/ends reasoning focuses attention on what I have elsewhere called (following Shelly Kagan[14]) *pure instrumentalism*—the thesis that reasons are communicated across causal, criterial, and mereological relations And this thesis *is* noncontroversial But neo-Humean instrumentalism is also defined by what I follow Stephen Darwall[15] in calling the *desire-based reasons* thesis reasons (in the sense relevant to the rational advisability of action and the appraisal of agents as rational) are grounded, ultimately, in the subjective, contingent, conative states of the agent—in her intrinsic valuings, for example [16] Of course, one can, and many do, accept pure instrumentalism without accepting the desire-based reasons thesis The desire-based reasons thesis is not an instrumental principle and so, neo-Humean instrumentalism (which includes both principles) "goes beyond" a purely instrumental principle

Roughly, the idea seems to be this Everyone should admit that a person has reason to undertake the means to those ends she has reason to bring about That is just the thesis of pure instrumentalism But the desire-based reasons thesis asserts that there is a reason for an agent to bring about the ends that she has (what she intrinsically values, for example) And this reason does not, and cannot, derive from these things being necessary means to our ends

---

[13] "Levi against U-Maximization," this JOURNAL, LXXX, 9 (September 1983) 531-34

[14] "On Instrumental Reasons," presented at the American Philosophical Association, Central Division Meeting, April 29, 1989

[15] *Impartial Reason* (Ithaca Cornell, 1983), p 27

[16] That an individual's intrinsic valuings are subjective, contingent, conative states of the agent is not completely uncontroversial I develop and defend the claim that they are such states in "Values and Desires" (typescript) While I favor a form of neo-Humean instrumentalism that grounds an agent's reasons on her intrinsic valuings, for present purposes this can be treated as a mere example Those who challenge the subjective, contingent, conative status of an individual's intrinsic valuings may substitute another, less controversial, mental state such as preference or desire

Korsgaard confuses the reader about her real complaint by suggesting that the feature she alludes to here arises because the neo-Humean makes a distinction between an agent's ends and what he is going to do If I am right that her concern is that the principle of pure instrumentalism is only about the transmission of reasons and some further normative principle is needed to ground instrumental reasoning, the "further normative principle" would be just as necessary for those who draw no such distinction Let us call this the *grounding problem*

VI QUINN'S QUESTION

Korsgaard is not alone in worrying about whether neo-Humean instrumentalism can yield reasons that are grounded in a way that is necessary if they are to be normative Warren Quinn (*op cit*) shares the concern

Suppose, with Quinn, we take neo-Humean instrumentalist theories to attempt to ground judgments of practical rationality ultimately on the agent's pro- and con-attitudes and that we understand these to be "functional states that dispose us to act " Quinn asks "whether pro- and con-attitudes conceived as functional states that dispose us to act have any power to rationalize those acts" (*op cit*, p 236) To answer this question, Quinn imagines a peculiar functional state that disposes its possessor to turn on radios she sees to be turned off And he says "I cannot see how this bizarre functional state in itself gives me even a prima facie reason to turn on radios, even those I can see to be available for cost-free on-turning" (*op cit*, p 237)

An answer to Quinn will wait Here, I want to accept his implicit criterion for a normative theory of practical rationality conformance with the principle must seem (at least on reflection) to rationalize the actions

Quinn's concern, remember, was how "pro- and con-attitudes conceived as functional states that dispose us to act have any power to rationalize those acts " He did not see how a mere disposition to perform an action could even tend to show that the agent had a reason to perform that action

I think there is a fairly obvious response to Quinn's concern It is this if pro- and con-attitudes are conceived only as functional states that dispose us to act, they do not have any power to rationalize those actions Quinn is right to say that a mere disposition to turn on radios in one's vicinity does not give one "even a prima facie reason to turn on radios " But a mere disposition to perform an action is clearly not constitutive of having a pro-attitude toward that action A person's disposition to fall asleep at department meetings, get tongue-tied when talking with a celebrity, or display anger when talking to an ex-spouse hardly shows that the person intrinsically values these actions or any state of affairs to which she believes them to be instrumental. For all Quinn's examples show, the problem is not that certain pro-attitudes cannot rationalize actions but, rather, that these pro-attitudes cannot be adequately understood as merely dispositions to act or be motivated to act—an analysis that, to my knowledge, has not been proffered by any defender of a neo-Humean instrumentalist position.

Interestingly, in setting up the problem—just a paragraph before he actually raises it—Quinn recognizes three separate times that the neo-Humean instrumentalist thinks of the functional state of disposing one to perform an action as only one (if a central one), among several, elements of a pro-attitude. And, so, the question naturally arises why Quinn thinks he has raised a difficult problem for the neo-Humean. Does not the question of whether a functional state that disposes us to act could have any power to rationalize the act obviously depend on what other characteristics that functional state has?

What follows is quite speculative, of course, but I suspect that Quinn thought that nothing one could plausibly add to the mere disposition to act in order to get a full description of a pro-attitude would do any more to rationalize the act. Infuse a pro-attitude with as much functional meaning as you want—inject into it as much qualitative content as you will—pro-attitudes will not rationalize actions. They cannot ground reasons for acting. For example, treat a pro-attitude toward some state of affairs not only as including a disposition to bring about that state of affairs, but also a disposition to feel regret at missed opportunities to do so and a positive affective mental state toward that state of affairs. We can still wonder whether there is anything in all of this that rationalizes the act in question—whether the attitude toward the state of affairs in question grounds a reason to bring it about.

If this is the thought that lies behind Quinn's concern, then I think it is just the worry Korsgaard has. Transmitting reasons from ends to means is noncontroversial. But how does the neo-Humean instrumentalist "get reasons going" in the first place? It must be by unwittingly and illicitly appealing to a noninstrumental normative principle to ground practical reason—something like "one ought, rationally, to bring about one's ends."

VII NORMATIVITY ON A NON-NORMATIVE FOUNDATION

The problem The worry, then, is fairly simply put The principle of pure instrumentalism ensures the transmission of reasons from ends to means, but the neo-Humean, no less than anyone else, needs a substantive and noninstrumental principle of rationality to give rational endorsement to the ends—to ground the chain of reasons No reasons will be transmitted to the means unless there is a reason for the ends This objection really has nothing to do with demandingness and the avoidance of normative triviality

Now, it might be thought, we have a puzzle for the neo-Humean instrumentalist (If I am right, this is really what lies behind the concerns about normative triviality) Suppose he makes the following claims

(a) Whatever is endorsed by practical reason is so endorsed by virtue of being instrumental to an agent's ends
(b) Ends cannot convey reasons unless there are reasons for the ends
(c) Ends are beyond intrinsic rational evaluation

He is committed, then, to the conclusion that nothing can be endorsed by practical reason If, on the other hand, he claims that practical reason endorses an agent's ends (no matter how defined), then, no less than any other theory of practical rationality, he owes us a justification of this clearly noninstrumental principle

How might the neo-Humean instrumentalist reply? My answer will begin far from theories of practical rationality, but not quite as far away as the story of *The Little Prince* took us

Excursus into legal philosophy Through the middle of the twentieth century, there was a debate among legal positivists Unlike their natural-law opponents, legal positivists saw law as being solely a product of human activity Natural-law theorists saw legal validity as requiring some sort of conformance of human dictates with the requirements of a natural law, which exists independently of human activity Positivists denied this Instead of subjecting statutes and orders to an external test of legal validity, they employed a "pedigree test" of legal validity a statute is valid law, for example, if it is validly derived from (or enacted in accordance with) a more basic valid law Of course, the pedigree cannot go back forever To avoid regress, this chain has to end in a fundamental law

Hans Kelsen[17] called this fundamental law the "*Grundnorm*" The *Grundnorm* exists because it is actually obeyed But existence is not validity So, the question arises, is this fundamental law valid or not?

---

[17] *The Pure Theory of Law* (Berkeley California UP, 1967)

If not, how could it confer validity on the "derived" laws? If so, from whence did its validity come? The pedigree test gives only a mechanism to "transmit" validity to subordinate norms To confer validity on the *Grundnorm* would be to "go beyond" the pedigree test But, if there could be an independent source of legal validity, why does it apply (only) to the fundamental norm that is actually effective in society? Why does this source not confer validity directly on each particular law? Why does it not confer validity on some other basic norm that satisfies a requirement different from social effectiveness? Why not to a norm that is morally demanded? Or to a norm that every rational being would will to be a universal law of nature insofar as reason had control of his will?

Kelsen's answer was that the validity of the *Grundnorm* is "presupposed" (*ibid*, pp 201-05)—an answer which, to quote Bertrand Russell[18] speaking of another matter, has all "the advantages of theft over honest toil" (*ibid*, p 71) Kelsen's "answer" addresses none of the concerns we just raised about the validity of the *Grundnorm*, and so, as Russell urges his readers, "Let us proceed with our honest toil" (*ibid*)

H L A Hart[19] shared Kelsen's legal positivism, but offered a different account of the validity of "derived" laws They did not have to receive their validity by being "derived" from a valid law—as if validity poured from the basic rules to the derived rules through the conduit of valid derivation Rather, the basic rule of law, which Hart called "the ultimate rule of recognition," defines "validity" within the legal system in question Given this role, the question of the validity of the ultimate rule of recognition does not arise It involves a sort of "category mistake" to think of the ultimate rule of recognition as being valid *or* invalid Hart says

> We need the word 'validity', and commonly only use it, to answer questions which arise *within* a system of rules where the status of a rule as a member of the system depends on its satisfying certain criteria provided by the rule of recognition No such question can arise as to the validity of the very rule of recognition which provides the criteria, it can neither be valid nor invalid but is simply accepted as appropriate for use in this way To express this simple fact by saying darkly that its validity is 'assumed but cannot be demonstrated', is like saying that we assume, but can never demonstrate, that the standard metre bar in Paris which is the ultimate test of the correctness of all measurement in metres, is itself correct (*ibid*, pp 105-06)

---

[18] *Introduction to Mathematical Logic* (London Allen and Unwin, 1919)

[19] *The Concept of Law* (New York Oxford, 1961)

The result is that rules of law may be valid in virtue of their relation to a fundamental rule, even though this fundamental rule is neither valid nor invalid Its status as law is secured by a social fact its general acceptance in a population, among other things Unlike other rules of law, the ultimate rule of recognition is not law because it is valid, it is law because it is accepted This is problematic if one thinks of the resultant validity, legal validity, as unconditional—if one thinks of the normativity of law as being unqualified A classical natural-law theorist, for example, will worry about how a fact about social acceptance can ground the validity of a norm of law

But, on Hart's view, laws are valid only *relative to a system of law* To call a statute, for example, valid law is to say that it is properly created within some (usually implicitly indicated) legal system This is what Kelsen misses He has only half digested the implications of legal positivism Recognizing that law is a product solely of human activity, he still looks for a concept of validity that can be applied from some external perspective so as not to make validity a "system bound" concept He "presupposes" the validity of the *Grundnorm* because he does not fully appreciate that all we seek to say when we talk of the validity of a law is its conformance with the basic norms of the legal system That is what "legal validity" means

Application to reasons for acting Now, what application do the lessons of this excursion into the recent history of legal philosophy have for the normativity of instrumental reasons and the viability of neo-Humean instrumentalism as a theory of practical rationality? As a transition, let us begin by talking about legal reasons for acting

When we say that a person has a legal reason to perform an action, we justify this by appeal to, for example, specific statutes or court orders that the person has a legal reason to obey If called on to justify the claim that the person has a legal reason to obey this specific statute or court order, we will cite more basic legal norms which, we argue, the person has a legal reason to obey and which endorse this statute or court decision But the (legal) reason giving comes to an end at some point And, in Hart's view, that point is at a complex social fact about rules that are in effect and the attitudes people have toward those rules

To be sure, this makes legal reasons much less impressive than they might otherwise seem One might very well recognize that he, or others, have a legal reason to do something but correctly think this conclusion is completely unimportant in deciding what to do And, this is part of the message of legal positivism, too Making legal validity and the existence of legal reasons depend on a logically and morally arbitrary social fact is unacceptable if we want claims of legal validity and what one has legal reasons to do to have all of the grandeur that the natural-law theorist invests in them But this is a reason to reject a hybrid theory that takes the account of legal validity and legal reasons from the positivist and the significance of legal validity and legal reasons from the natural-law theorist It is not a mark against the legal positivist

Finally, now, to apply these insights to the issue of the normativity of instrumental reason and the adequacy of neo-Humean instrumentalism, let me begin by noting a rich and important dissimilarity between the two issues Recall that Hart says "we need the word 'validity', and commonly only use it, to answer questions which arise *within* a system of rules " We can say no such thing about the concept of a reason for acting Indeed, part of the problem with reconciling the competing theories of reasons for acting is that we have in mind too many different ideas and too many different uses for the concept[20] But we can say this one thing we seek to do when talking about reasons for an agent to act is to evaluate actions from within the

---

[20] When we talk about reasons for acting, we are often concerned with the actual reasoning process employed by the agent—not the issue of what action is rationally advisable from any standpoint This is, of course, parallel to discussions of theoretical rationality Our concerns are often with the agent's belief-forming processes and the evaluation of an agent's dispositions to believe—not with the question of what beliefs are rationally recommended Similarly, when we talk about reasons for acting, we are often concerned with the agent's intention-forming processes and the evaluation of the agent's dispositions to act

But, even where we are concerned with the evaluation of the actions themselves, we have a variety of interests At times, our interest is in endorsing an action *sans phrase* With respect to such evaluation, I see no reason to assume the evaluative standpoint of the agent We are also concerned, however, to evaluate actions from the evaluative standpoint of the agent Even when we do this, we have a variety of concerns

Sometimes we are concerned to evaluate actions in relation to the agent's desires (preferences, values, or some such state) and the agent's actual beliefs We want to know if the action is the one that would be rationally advisable if the world were as the agent believed it to be—regardless of whether the agent's beliefs were true, or even reasonable An agent who performs actions that would not achieve her ends were the world as she believes it to be makes a special kind of practical error

Sometimes we are concerned to evaluate actions in relation to the agent's desires (preferences, values, and the like) and the beliefs it would be reasonable for the agent to have Here, we seek to determine whether the action was rationally advisable given the epistemic situation in which the agent found herself Agents who perform actions that would not achieve their ends were the world the way it is rational for them to believe it to be, may be making the same practical mistake noted above, or they may be making an epistemic mistake that has practical implications for them

Sometimes, I contend, we are concerned to evaluate action in relation to the agent's desires (preferences, values, and the like) and the actual state of the world Here we are concerned with the advisability of the action from the agent's evaluative standpoint This is what I call *rational advisability* agent's system of norms—in relation to what the agent intrinsically values, one thing we seek to do when we talk about the practical rationality of an agent is to evaluate the coherence of her pursuits and intentions from within the agent's system of norms—in relation to the agent's values. It is this sense of there being a reason for an agent to perform an action (of the action being rationally advisable), and this sense of an agent being practically rational, that the neo-Humean instrumentalist seeks to explicate.

Of course, the neo-Humean instrumentalist must engage in the same "lowering of expectations" that the legal positivist must. Saying that there is a reason in this sense for an agent to perform an action does not mean anything more than that this action best promotes the agent's ends. It certainly does not, by itself, support what Allan Gibbard[21] has called "a flat, flavorless endorsement" (*ibid.*, p. 7).

Ought people to take efficient means to their ends? If the 'ought' is one of rational advisability, the neo-Humean instrumentalist says, then, (trivially) they ought. If the 'ought' is an expression of the speaker's acceptance of norms that recommend his performing the action, there is no reason to think that he ought. I would, for example, urge those whose fundamental values are those of the National Socialist Party to act contrary to the reasons they have in the neo-Humean instrumentalist sense of reasons.

VIII. NEO-HUMEANISM UNDERSTOOD ARIGHT

Neo-Humeans do not believe—or, at any rate, ought not to believe—that practical rationality is an objective matter. To be sure, there may be a fact of the matter about what action is rationally recommended for an agent. This fact, though, depends crucially on subjective states of that agent. On my view, it is the agent's values that play this role in determining what is rationally advisable for the agent. Other neo-Humeans focus on desires or preferences. This is an internecine dispute; it is not important here. All neo-Humeans take the rational advisability to depend, ultimately, on subjective, contingent, conative states of the agent in questions.

But no neo-Humean should think that, at bottom, actions are rationally advisable because they are instruments to rationally advisable ends (aims, projects, the satisfaction of desires or preferences, or the promotion of values). Certainly, actions can be rationally advisable for this reason, but this cannot be what rational advisability is at bottom. For at bottom, there is a fact—a brute fact—about the agent's subjective, contingent, conative states And the critics of neo-Humeanism are right to say that, on the neo-Humean view, this fact is not intrinsically rationally appraisable Their mistake, the neo-Humean contends, is in thinking that it must be

---

[21] *Wise Choices, Apt Feelings: A Theory of Normative Judgment* (New York: Oxford, 1990).

Critics of the neo-Humean position have implicitly accepted a view of rational advisability which is analogous to the view of legal validity which was accepted by critics of legal positivism This view holds that, while an action's rational advisability might depend on the action having a proper relation to some end or goal, this relationship could confer rational advisability on the action only if the end or goal were rationally advisable Similarly, the critic of legal positivism holds that, while a rule's legal validity might well be dependent on its being validly enacted in accordance with some more basic legal rule, this relationship could confer validity only if the more basic rule were, itself, legally valid At first blush, this seems eminently plausible—indeed, almost inescapable But escape it we must if we are to be neo-Humeans

Kelsen, for all his brilliance in developing the modern form of legal positivism, could not shake the intuition that the basic rule of law had to be legally valid in order for other legal rules to be legally valid in virtue of some relation to it Because his conception of legal validity was still clouded by intuitions based on a natural-law tradition, he was forced to resort to presupposing the legal validity of the basic rule of law The Kelsenian picture must be rejected if we are to understand neo-Humeanism aright

The neo-Humean adopts a view of how an agent's fundamental ends determine the rational advisability of her actions that is more similar to Hart's view of how the basic legal rule determines the legal validity of other rules of a legal system For the neo-Humean, the agent's ultimate ends (I would say, "her intrinsic values") are neither rationally advisable nor rationally inadvisable, in themselves They are, rather, the brute facts about the agent's psychology in virtue of a relationship to which policies, plans, and actions can be rationally advisable or inadvisable The agent's ultimate ends confer this status on policies, plans, and actions not because these ends have some normative standing in themselves They do this because the particular property of being rationally advisable just is the property of being properly related to these brute facts (just as the property of being legally required just is the property of being properly related to some complex brute fact about a society)

The theory of rational advisability offered by the neo-Humean does not give rational advisability all the status some would want in such a concept It certainly does not proffer a simple, unqualified answer to the question of what to do That is fine We advise people from many different perspectives Sometimes our advice explicitly indicates the perspective, sometimes the indication is contextual Sometimes our advice is based on norms that we, ourselves, adopt, sometimes it is based on norms we apply without endorsement Rational advisability is advisability from just one perspective Because this perspective is constructed from the agent's own values, it is one to which the agent has a special relationship[22] Still, there are, as the critic of the neo-Humean position wants to insist, many ways in which we might want to evaluate actions other than from the normative perspective of the agent The neo-Humean should not deny this All she must deny is that these other sorts of evaluation constitute evaluations of the rational advisability of actions

IX CONCLUSION

On the neo-Humean's view, judgments of rational advisability do not address the question of how to act, simpliciter Still, they address questions which are important and in which we are interested They address the questions of what actions are recommended by an agent's ends and when an agent has the intentions and pursuits appropriate to her ends One is free to think these questions uninteresting Disinterest cannot be refuted But, for those interested in these questions, the neo-Humean instrumentalist offers answers that do not require anything "beyond" instrumentalism It accomplishes this by construing instrumental rationality as having a sort of "groundless normativity "

DONALD C HUBIN

Ohio State University

---

[22] I defend the claim that the neo-Humean can defend a special motivational status for claims of rational advisability in "What's Special about Humeanism," *Noûs*, xxxiii, 1 (1999) 30-45

OF COLORS, KESTRELS, CATERPILLARS, AND LEAVES*

According to color realism, object colors are mind-independent properties that cover surfaces or permeate volumes of objects In recent years, some color scientists and a growing number of philosophers have opposed this view on the grounds that realism about color cannot accommodate the apparent unitary/binary structure of the hues For example, C L Hardin[1] asserts

> the unitary-binary structure of the colors as we experience them corresponds to no known physical structure lying outside nervous systems that is causally involved in the perception of color This makes it very difficult to subscribe to a color realism that is supposed to be about red, green, blue, black, and white—that is, the colors with which we are perceptually acquainted (*ibid*, p 300)

Similarly, Evan Thompson[2] says

> There is no distal-based scheme that can generate the basic chromatic categories of human color vision (red, green, yellow, and blue) plus the resemblance and difference relations among hues that result from the opponent structure of these categories (red-green, yellow-blue) (*ibid*, p 9)

And Leo Hurvich,[3] one of the founders (along with Dorothea Jameson) of the opponent-processing theory in color science, comments

> it should be clear by now that object color is not something that inheres in objects (*ibid*, p 61)

Another related threat is posed by the possibility of higher-dimensional color distinctions in other animal species While most humans are trichromats, there is strong evidence to suggest that some fish, birds, and turtles are tetrachromats, and pigeons may be pentachromats This evidence compounds the problem posed by the unitary/binary structure of the hues, since it may well be that tetra-

---

* Talks based on this paper were given at Duke University and the University of Athens We would like to thank Alex Byrne for his written comments, and Stuart Church for the use of his images, reproduced here as figure 1

[1] "Reinverting the Spectrum," in Alex Byrne and David Hilbert, eds , *Readings on Color*, Volume 1 (Cambridge MIT, 1997), pp 289-302

[2] "Colour Vision, Evolution, and Perceptual Content," *Synthese*, civ (1995) 1-32

[3] *Color Vision* (Sunderland, MA Sinaeur, 1981)

0022-362X/01/9809/469-87

chromatic vision involves not just unitary and binary hues, but ternary as well [4]

Thompson, Adrian Palacios, and Francisco J Varela put the general difficulty for the view that colors are external, physical properties in the following way

(1) For something to be a chromatic color it must be a hue
(2) For something to be a hue it must be either unique or binary (or ternary)
(3) Therefore, if hues are to be reductively identified with physical properties, these physical properties must admit of corresponding unique, binary (or ternary) divisions
(4) Organism-independent, external properties such as light waves and spectral reflectances do not admit of such divisions
(5) Therefore, color cannot be reductively identified with such organism-independent, external properties (*ibid*, p 15)

This argument, of course, is an argument for external irreducibility It does not yet undermine color realism For colors might be mind-independent, external properties that supervene upon physical properties without themselves having a hidden physical nature Alternatively, colors might be emergent properties without a physical supervenience base, properties whose natures are wholly given to us in sense experience Neither of these options is at all attractive, however The former must countenance implausible epistemically basic yet metaphysically derivative bridge laws that link colors with underlying physical properties [5] The latter has the consequence that colors are causally inefficacious with respect to the physical world For if colors emerged from the underlying physical configurations without supervening upon them, then the colors of things might have been different while the underlying physical facts and laws remained the same In that case, which colors objects have (or indeed whether they have any colors at all) makes no difference to their physical interactions And that intuitively is false Painting the walls of a room yellow causes it to be brighter than would painting the walls brown Staring at a bright red light and looking away causes a characteristic neural response that generates a green afterimage Moreover, if colors make no difference to how light is reflected from objects, to the subsequent changes at the retina, in the optic nerve, and so on, then it follows that we do not see colors?[6] That seems absurd Worse still, if we do not see colors, then intuitively we do not see things at all Intuitively, we see the facing surface of things by seeing their colors

---

[4] See Thompson, Adrian Palacios, and Francisco J Varela, "Ways of Coloring Comparative Color Vision as a Case Study for Cognitive Science," *Behavioral and Brain Sciences*, xv (1992) 1-74

[5] For further criticism, see Tye, *Ten Problems of Consciousness* (Cambridge MIT, 1995), chapter 2

So it seems that, if colors are external, mind-independent properties, then they must be reductively identifiable with physical properties But how can such a reductive identification preserve the unitary/binary division of the hues and relatedly the possibility of higher-dimensional hue divisions elsewhere in the animal kingdom? We shall offer an answer to this question here In our view, the recent tendency to suppose that we mistakenly project colors onto the world—their real home being in the mind and in the opponent structure of the relevant neural processing—is without adequate foundation Colors, we maintain, are objective properties, out there in the natural world That view, we shall argue, is perfectly compatible with color science

The paper is divided into four sections In section I, we make some clarificatory remarks about the unitary/binary distinction and present some reasons from scientific research on comparative color vision for taking seriously the claim that there may well be colors—unitary, binary, and ternary—humans cannot see In section II, we show how a simplified version of the opponent-processing theory in color science is supposed to provide a subjectivist explanation of the division of humanly perceptible hues into unitary and binary, and also how this explanation may be extended straightforwardly to ternary combinations of hues In section III, we offer an alternative explanation of the divisions among the hues, which locates these divisions out in the world In section IV, we consider an objection to realism for humanly inaccessible hues, based upon a case from comparative vision studies As will be seen, our approach is strongly influenced by the opponent-processing theory of color vision, but it resists the subjectivist metaphysical implications many adherents of the theory have mistakenly read into it

I ON HUES, UNITARY, BINARY, AND TERNARY

Some hues are unitary in particular, red, yellow, green, blue Others are binary for example, orange, purple, lime Binary colors are always mixes of other colors Orange is reddish yellow, lime is yellowish green, purple is bluish red Unitary colors have shades that are not mixes of other colors Red, for example, has a shade that is not a combination of any other hues This is unique red

---

[6] We assume here that seeing *x* demands a causal connection with *x*

What is meant by saying that binary hues are mixes or combinations, whereas unitary hues are not? One reading of this claim is in terms of perceptual mixing.[7] Something, in looking orange, for example, looks a bit reddish and a bit yellowish, but there is a shade of red such that anything in looking that shade does not look another color at all. Intuitively, however, when we say that orange is reddish yellow, we are not merely making a remark about the look of things. Intuitively, our supposition is that orange is a color such that something, in being orange, *is* a bit reddish and it *is* also a bit yellowish; moreover, any given object, in being a bit reddish and a bit yellowish in the same surface regions, *is* orange in those regions. Orange, thus, is naturally taken to be a conjunctive property: reddishness and yellowishness.[8] That, we suggest, is the common-sense view. And that provides the basis for the most straightforward and direct way of understanding the unitary/binary distinction. Binary chromatic colors are conjunctive properties whose conjuncts are exactly two of the following: reddishness, yellowishness, bluishness, greenishness. Unitary chromatic colors are colors for each of which there is a shade that is not conjunctive in this way.

The unitary/binary distinction, as we have just stated it, applies only to the hues humans can see. But there is evidence from scientific work on comparative color vision, that some hues are humanly inaccessible. Once these hues are admitted, a case can be made for supposing that they admit of corresponding unitary and binary classifications, and perhaps even ternary classifications, too. Consider first some facts about kestrel vision.

The vole population in Northern Europe oscillates in a four-year cycle. A crash in the vole population in a region can result in a thousand kilometer migration of those kestrels which feed primarily on the voles. Up until a few years ago, we knew almost nothing about how the kestrels are able to track the vole population over such large areas of land. J. Viitala, E. Korpimaki, P. Palokangas, and M. Koivula[9] recently discovered, however, that when we illuminate vole scent marks (which male voles use to indicate their paths) with light from the ultraviolet (UV) portion of the spectrum, the marks contrast with their background. To normal human perceivers in standard conditions, these scent marks are indistinguishable from their back-grounds To kestrels, however, they are not Through extensive field experiments on wild kestrels, Viitala, Korpimaki, Palokangas, and Koivula showed that under normal lighting conditions, kestrels concentrated their hunting in areas containing straw that had been soaked in vole feces and urine, while ignoring areas containing straw that had been soaked in water Prima facie, then, kestrels are seeing a color of vole urine we humans do not see—a color, the detection of which requires sensitivity to UV light[10]

---

[7] See Hardin, p. 291.

[8] This was suggested by Byrne and Hilbert, "Colors and Reflectances," in Byrne and Hilbert, eds., pp. 263-88, here p. 280.

[9] "Attraction of Kestrels to Vole Scent Marks in Visible in Ultraviolet Light," *Nature*, CCCLXXIII (1995): 425-27.

Behavioral evidence supporting the possibility of UV vision in birds is not limited to raptors Many of the thirty-one species of birds, whose UV vision has been demonstrated since 1972, forage on insects[11] As we might expect, many of those insects reflect light strongly in the UV range

The caterpillars of the gray shoulder knot moth (lithophane ornithopus) reflect maximally in the UV range In the human-visible portion of the spectrum, they reflect a large percentage of the incident light in the short wave-length region, and more in the middle than in the long Furthermore, gray shoulder knot caterpillars are typically found on oak leaves in Northern Europe, and oak leaves reflect only minimal percentages of UV light, but a large percentage of the incident light in the short wave-length region, and more in the middle than in the long To humans, a caterpillar sitting on an oak leaf looks to be very similar in color to its background But, to birds, the gray shoulder knot caterpillar contrasts strongly against the oak leaf Hence, we can hypothesize that the strong UV markings on the caterpillar fill the same role for birds that the bright red markings on the poison dart frog fill for humans they advertise distastefulness[12]

---

[10] One should not be tempted into thinking that the color birds see when looking at an object that reflects UV light is simply what we see when the object is illuminated by UV light alone Under only UV-wave-length light, the vole urine will look brighter to us, but we shall not see the hue the kestrel sees in that light (or in daylight)

Here is an analogous case Suppose we take an object that reflects only light from the long (red) portion of the spectrum and we place it on a background that reflects only light from the short (blue) portion of the spectrum Suppose we then illuminate the scene with light from only the long (red) portion of the spectrum The object will reflect light, but the background will not Now suppose we let a true color-blind person (monochromat) view the scene Monochromats certainly can discriminate objects that reflect light from objects that cannot, and so to the monochromat the object will look brighter than the background But evidently it would be a mistake to infer from this that the monochromat is seeing the color trichromats see as they view the object in the same circumstances (or in daylight)

[11] For an overview, see A T D Bennett and I C Cuthill, "Ultraviolet Vision in Birds What Is Its Function?" *Vision Research*, xxxiv (1994) 1471-78

[12] Cuthill, J C Partridge, Bennett, S C Church, N S Hunt, and S Hunt, "Ultraviolet Vision in Birds," *Advances in the Study of Behavior*, xxix (2000) 159-214

Prima facie, thus, birds detect a color on the caterpillar which humans do not.

The evidence for higher-dimensional color vision in birds is not restricted to observation of behavior. There is also physiological data. In recent years, biologists have isolated at least four spectrally distinct cone types in birds compared with our three. For example, four have been discovered in the retinas of mallards and five in the retinas of pigeons. The maximal sensitivity of avian cones is in the near UV range, varying from 355nm in the Pekin robin (or red billed leiothrix) to 403nm in the penguin.[13] Furthermore, avian cones contain droplets of oil that absorb parts of the 300-700nm range of light. Numerous different kinds of oil droplets have been identified, signifying a wide variety of color vision across avian species.[14]

It may be tempting to insist that birds cannot see colors we humans do not see, even though they do detect qualities that play a comparable role with respect to the birds' visual systems and behavior. For colors are fixed by their locations in color space, and there is simply no room in our color space for a region occupied by a genuinely novel hue.

This last remark is true, but it is far from decisive. Our color space is indeed completely filled. Its geometry cannot allow a completely new hue.[15] But that is compatible with supposing that there are novel color spaces with different geometries. For example, there might be a tetrachromatic space, closed with respect to hue-saturation-lightness relations, which is connected by routes of resemblance to our color space and which contains the latter space as a part.[16] Parts of this space would be closed off from human trichromats in much the way that parts of our color space are closed off from human dichromats (pure red-green dichromats see only yellows and blues among the chromatic colors). So, we would be unable to form a clear conception of what it is like to experience some of the colors in this tetrachromatic space, just as a pure red-green dichromat cannot form a clear conception of what it is like to experience red

---

[13] While 403nm is within the human visual spectrum, the sensitivity of the relevant cone in the penguin's visual system extends further (into the UV range).

[14] For overview, see Bennett and Cuthill.

[15] Sensitivity to hue, saturation, and brightness varies among people. It may be the case that, while keeping saturation and brightness fixed, Joe can discriminate more red hues than can Bob. But any red hue Joe can discriminate lies between red hues Bob can discriminate. So, each of the former hues has a place within Bob's color space in between hues Bob can pick out. Such hues, therefore, are not genuinely novel. Our color space already includes all hues than can possibly lie between the hues we know, and therefore it is closed with respect to hue.

[16] This suggestion was made by Thompson, "Novel Colors," *Philosophical Studies*, LXVIII (1992): 321-49.

There is no general agreement as to just which features must be shared by *all* alternative color spaces It seems reasonable, however, to hold that a sufficient condition for qualities to count as hues is that they be discriminated via the eyes in response to the reflected light, that they exhibit color constancy[17] and color contrast effects, and that they belong to a family of qualities with these effects (a family whose members are connected to one another by routes of resemblance), whether or not they are accessible to human vision [18] Color constancy and color contrast effects have been found manifested in the behavior of birds, goldfish, and honeybees [19]

Color contrast is the product of opponency among the hues In our color space, red is opposed to green, yellow to blue This is reflected in our color experience We cannot conceive of experiencing a surface that is both red and green all over at the same time Likewise for blue and yellow If opponency of this sort is essential to something's being a hue, and it certainly seems essential to our hues, then humanly inaccessible hues will require opponent complements So, if there is a hue we cannot experience, there must be two—that hue and its opponent complement Accordingly, any novel tetrachromatic color space will need at least one opponent pair of novel hues among its six primary hues Moreover, since our color space includes binary combinations of primary hues (orange (reddish yellow), lime (greenish yellow), purple (reddish blue), turquoise (bluish green)), there is some reason to think that novel tetrachromatic spaces will have binary combinations involving the novel hues and arguably even ternary combinations too, given the three opposing pairs of primary hues

On this understanding of tetrachromatic vision, tetrachromats are not only able to see hues that we trichromats are unable to see, but they are (or may well) be able to see combinations of these hues which are inaccessible to us The problem of accounting for the distinction between unitary and binary hues now becomes a special case of a more general problem, that of providing an objective, external ground for the distinction between what might be called

---

[17] For remarks on color constancy, see below, p 479

[18] We address the question of what it is for a quality to be a color in detail in our "In Defense of Color Liberalism" (in preparation)

[19] For an overview, see Christa Neumeyer, "Comparative Aspects of Color Constancy," in Vincent Walsh and Janusz Kulikowski, eds , *Perceptual Constancy Why Things Look as They Do* (New York Cambridge, 1998), pp 323-51

basic and *derivative* hues And the prevailing view is that this problem cannot be solved by the color realist even in the unitary/binary case We disagree

II A SIMPLIFIED QUANTITATIVE MODEL OF OPPONENT PROCESSING AND SUBJECTIVISM ABOUT THE HUES

The model we shall consider initially is that presented by Hardin in *Color for Philosophers*[20] According to Hardin, chromatic color experience is generated by neuronal activity in two channels, one for red-green color experience and the other for yellow-blue Where '0' represents the base level of neural activity, that is, the level corresponding to no visual response, and 'L', 'M', and 'S' represent the neural activity in the relevant neurons connecting to the long, medium, and short wave-length cones in the eye, the difference in L and M activity yields the red-green channel while the sum of L and M activity minus the S activity gives the yellow-blue channel Specifically, where $L - M$ is the red-green signal,

$(L - M) > 0$ yields the experience of (pure) red

and

$(L - M) < 0$ yields the experience of (pure) green

assuming the yellow-blue channel is in balance Similarly, where $(L + M) - S$ is the yellow-blue signal,

$((L + M) - S) > 0$ yields the experience of (pure) yellow

and

$((L + M) - S) < 0$ yields the experience of (pure) blue

assuming the red-green channel is in balance

This model is oversimplified, as Hardin notes, since (at a minimum) the above formulae should be assigned coefficients that correspond to appropriate weightings for the three cone types (corresponding to the proportions of each cone type in a given retinal region, the relative efficiency of each cone type in generating an output from the radiant energy, and so on)[21] But it suffices for present purposes

How is this model relevant to the unitary/binary distinction among the hues? Recall that, for Hardin, the claim that orange is reddish yellow is to be understood in terms of perceptual mixing Orange supposedly is a perceptual mix of red and yellow On this view, one who has the experience of red and who also the experience of yellow thereby has the experience of orange What is required, then, for the experience of orange, on the above model, is that the neural activity in the red-green channel be greater than zero and that the same be true for the activity in the yellow-blue channel This, according to Hardin, is what grounds the claim that orange is a binary hue By contrast, pure red is a unitary hue only in virtue of the fact that the experience of pure red occurs if and only if (i) the neural activity on the red-green channel (L – M) is greater than zero (thereby precluding the experience of green) and (ii) the neural activity on the yellow-blue channel ((L + M) – S) is in balance (thereby precluding the experience of either blue or yellow)

---

[20] Indianapolis Hackett, 1988

[21] For more on qualifications to the simple model, see below, p 482

On this account, the distinction between unitary and binary hues is grounded upon facts about neural processing It is also sometimes claimed that the hues themselves are best viewed as features of that processing—features that color experiences mistakenly represent as belonging to external things According to this view, evolution has seen fit to wire us so that we project colors onto the world, but their real home is in the brain Two well-known cognitive scientists, Leda Cosmides and John Tooby, in their introduction to Simon Baron-Cohen's recent book *Mindblindness*,[22] put the point this way

> The machinery that causes [color] experiences allows us to identify something as the same object across situations despite the different wavelength composites that it reflects from circumstance to circumstance Far from being a physical property of objects, color is a mental property—a useful invention that specialized circuitry computes in our minds and then "projects onto" our percepts of physically colorless objects This invention allows us to identify and interact with objects and the world far more richly than we otherwise could That objects seem to be colored is an invention of natural selection, which built into some species, including our own, the specialized neural circuitry involved (*ibid*, p xi)

Turning now to tetrachromatic vision, there is empirical evidence, as we noted in the last section, that tetrachromats see hues that trichromats cannot[23] In order to extend the opponent-processing theory into tetrachromatic color space, it is necessary to hypothesize three chromatic channels instead of two. Further, behavioral responses to UV stimuli require UV-sensitive channels. Oversimplifying greatly, the three chromatic channels proposed by Thompson, Palacios, and Varela (*op. cit.*, p. 470) for the goldfish are as follows:

---

[22] Cambridge MIT, 1995

[23] It is worth pointing out that, even for trichromats, the four unitary hues need not always be red, green, blue, and yellow, as they are for human beings It may be that, in some cases, the responses of the opponent processing channels are shifted along the spectrum For example, honeybees are trichromats, but their vision is sensitive to UV light R Menzel and W Backhaus—"Color Vision in Honey Bees

$$C1 = (UV + S) - L$$
$$C2 = UV + (S - M) + L$$

and

$$C3 = (UV - S) + (M - L)$$[24]

When activity on C1 is greater than zero, the experience of a novel hue (H^+) results.[25] When activity on C1 is less than zero, the experience of another hue (H^-) opposing H^+ occurs. Thus, just as it is impossible for trichromats to experience a binary combination of red and green, so too it is impossible for tetrachromats to experience a binary combination of H^+ and H^-.

According to the extended theory, tetrachromats can experience other binary combinations, however. Activity on the C1 channel, in combination with activity on one of the other two chromatic channels with the alternate channel in balance, will produce the experience of a novel binary hue: $H^+ + I^+$, for example, (where I^+ is a novel hue associated with C2, let us suppose). Furthermore, with both of the other two chromatic channels active (not in balance), the extended theory predicts the experience of novel ternary hues.

Phenomena and Physiological Mechanisms," in D. G. Stavenga and R. C. Hardie, eds., *Facets of Vision* (New York: Springer, 1989), pp. 281-97—hypothesize that honeybees have two chromatic channels $X = (UV - (B + G))$ and $Y = ((UV - B) + G)$, where B and G represent activity on neurons connected to cones that are maximally sensitive at 435nm and 540nm respectively. X and Y yield opposing hues according to activity on the channels. When activity on X channel is greater than zero, it yields a hue X^+, and when activity is less than zero, it yields X^-. The same holds for Y channel. Thus, for honeybees, there are four unitary hues (X^+, X^-, Y^+, and Y^-). Since the experience of each hue involves a response to UV light, and none of our color experiences does, it follows that honeybee color vision is truly inaccessible to humans.

[24] All three channels have zero crossings (indicating "rest" positions for neural activity) in the middle wave-length range. As the goldfish lives in water, we would expect that the goldfish's visual system would not be very responsive to light in the blue-green portion of the spectrum.

[25] It is important to note that, while all three channels will be responsive to ultra-short wave-length incident light, none of the hues associated with them should presumptively be identified with the hue, UV. The chromatic channel that yields the experience of blue in humans ($(L + M) - S$) responds to long wave-length incident light, yet we do not take that to show that the experienced hue is red.

The extension of the subjectivist approach to color for the tetrachromatic case is obvious Hue H^+ is really a feature of the neural processing of creatures able to experience H^+ The experiences of these creatures represent H^+ as out in the world, but nothing outside the creatures' heads has H^+ The binary hue, $H^+ + I^+$, is a combination of H^+ and I^+ *only* in that the experience of the first hue involves the experience of H^+ and the experience of I^+ Experience of $H^+ + I^+$ is produced via activity greater than zero on both the C1 and the C2 channels Ternary hues involve combinations of three unitary hues but only in a correspondingly subjectivist way

III COLOR OBJECTIVISM REGAINED

The obvious view of color, at least as far as common sense goes, is that the colors we see objects and surfaces to have are external properties of those objects and surfaces We think of colors as inhering in surfaces, as, for example, in the case of the red that covers the outside of my car We also think of colors as sometimes attaching to volumes, as in the case of the green that fills a piece of transparent green glass [26] We take it for granted that objects typically retain their colors when they are not seen, thereby helping us to re-identify them

Another important fact about color, which is manifest to us in our everyday life, is color constancy Objects do not typically appear to change their colors during the day as the sunlight changes Grass in the early morning looks to have the same color as it does at midday or late in the afternoon, even though the light is very different Nor does it make much difference to the perceived colors of objects—plants, for example—when they are moved from outdoors to a setting of illumination by incandescent lamps Moreover, wearing sunglasses has little effect on the colors objects appear to have [27]

Constancy in apparent color, we might add, is sufficiently robust that two objects with different colors can continue to appear as they normally do even when they reflect light to the eye of the very same spectral composition For example, a bluish object lit only by reddish light of the setting sun continues to look blue, and a reddish object illuminated by the bluish light of the sky continues to look red even though the light the objects reflect back to the eye is the same [28]

---

[26] Further, we allow that colors can belong to thin films, as in the case of the blue of the sky

[27] See Hilbert, *Color and Color Perception A Study in Anthropocentric Realism* (Stanford CSLI, 1987)

[28] Roger Shepard "The Perceptual Organization of Colors An Adaptation to Regularities of the Terrestrial World," in Byrne and Hilbert, eds , *Readings on Color,* Volume 2 (Cambridge MIT, 1997), pp 311-56

The fact that objects appear to retain the same color through a wide variety of changes in the illumination conditions (though certainly not all) strongly suggests that colors are illumination-independent properties of those objects But if this is the case, with which such external physical properties are colors to be identified? And how does that identification preserve an objectivist division of the hues into unitary, binary, ternary, and so on? Evidently, the disposition to reflect light of so-and-so wave lengths is not a suitable candidate for surface color For that varies greatly with the illumination conditions, and it also varies greatly among objects of the same color

The most promising objectivist proposal for surface color appeals to surface reflectance[29] The phenomenon of metamerism presents an immediate difficulty for this proposal, however Metamers are objects which appear the same color even to normal perceivers in daylight but which have different overall reflectances This shows that colors are not reflectances simpliciter But it leaves open the possibility that surfaces of the same color have reflectances with something important in common in virtue of which the surfaces have that color

Consider again the opponent-processing theory Recall that, according to the theory (in much simplified quantitative form), where $L - M$ is the red-green signal,

$(L - M) > 0$ yields the experience of (pure) red

and

$(L - M) < 0$ yields the experience of (pure) green

provided that the yellow-blue channel is in balance Similarly, where $(L + M) - S$ is the yellow-blue signal,

$((L + M) - S) > 0$ yields the experience of (pure) yellow

and

---

[29] The reflectance of a surface is given by stating, for each wave length in the spectrum visible to humans, birds, fish, and turtles (300 - 700 nm), the percentage of light the surface reflects at that wave length Proposals for color that appeal to surface reflectance are to be found in Hilbert, Mohan Matthen, "Biological Functions and Perceptual Content," this JOURNAL, LXXXV, 1 (January 1988) 5-27, Richard Grandy, "A Modern Inquiry into the Physical Reality of Colors," in David Weissbord, ed , *Mind, Value, and Culture Essays in Honor of E M Adams* (Atascadero, CA Ridgeview, 1989), pp 229-45, Tye, *Ten Problems of Consciousness,* and *Consciousness, Color, and Content* (Cambridge MIT, 2000), and Allan Gibbard, "Visible Properties of Human Interest Only," in Enrique Villanueva, ed , *Philosophical Issues Perception* (Atascadero, CA Ridgeview, 1996), pp 199-208

$((L + M) - S) < 0$ yields the experience of (pure) blue

provided that the red-green channel is in balance

Now assume, as is surely intuitively reasonable, that, ceteris paribus, our color experiences under normal viewing conditions are veridical It follows that, ceteris paribus, under normal viewing conditions, when we undergo opponent processing of a sort appropriate to the experience of red, something red is present[30] That processing, on the present proposal, is triggered by a key feature of the surface reflectances of red objects to which our visual systems are tuned Specifically, *given the assumption that the oversimplified opponent-processing model is correct,* where L^* is the amount of long wave-length light, M^* is the amount of middle wave-length light, and S^* is the amount of low wave-length light (corresponding to the neural activity, L, M, and S respectively), our claim is that a surface is (pure) red if and only if it has a surface reflectance that disposes it, under normal viewing conditions, to reflect light such that, for the reflected light, $M^* <_s$ (is significantly less than) L^* and $S^* =_a$ (is approximately the same as) the total quantity, $M^* + L^*$[31] (Pure) red, thus, in our view, is an objectivist disposition common to surfaces with different reflectances This makes surface redness appropriately illumination independent, for the disposition to reflect certain relative amounts of light given certain lighting conditions is fixed by the surface reflectance Parallel proposals can be constructed for the other humanly accessible hues[32]

It is worth emphasizing that no claim is being made that there is, in nature apart from us, anything special about the relevant surface dispositions *We* find them special because of how our visual systems are constructed[33] In this sense, the colors we see are indeed anthropocentric But their being so does not make them subjective They are, we claim, real, objective properties, even if they are of no particular interest to creatures lacking our visual apparatus The colors we see are tailored to the color detection system evolution has given us The two fit one another like hand and glove

It is also very important to appreciate that any actual counterexample to the proposal we have made will also be a counterexample to any subjectivist view of color based on the opponent-processing model presented earlier Qualify the latter appropriately to handle the counterexample—it is, after all, by admission, very oversimplified—and, we maintain, the correspondingly qualified version of the former will handle the counterexample, too This point needs developing further

---

[30] The ceteris paribus qualification is included here primarily because of contrast effects These experiential effects interfere with or mask the real color

[31] In their "Colors and Reflectances," Byrne and Hilbert make a remark about the color green which is in keeping with our proposal about red (p 265) But their remark is not developed into a general view, and it is not applied to the unitary/binary distinction

[32] We ignore light emitters here, but the extension of our view to emitters is relatively straightforward

[33] See Gibbard

Suppose, to begin with, that the formulae in the opponent-processing model are qualified by numerical coefficients assigned to the values of S, M, and L—coefficients that reflect the different proportions of cone types on the retina Patently, this poses no difficulty for our proposal, either quantitatively or conceptually Appropriate coefficients now need to be introduced for the values of S*, M*, and L* For example, if *q* and *r* are the coefficients assigned to L and M, then given that the ratio of long wave-length cones to medium wave-length cones is *q* to *r*, the revised requirement on L* and M* is that *q*L* be (significantly) greater than *r*M*

Suppose now it is insisted that no mere assignment of numerical coefficients to the formulae in the simplified model will give a fully accurate account of how the opponent-processing channels work The spectral sensitivities of the cones vary in addition to the proportion of each cone type, so that in reality, the relevant functions are nonlinear [34] This is the view taken by R W G Hunt,[35] for example According to Hunt, to generate satisfactory opponent channels, we need to take the square roots of S, M, and L before summing or subtracting We also need to qualify with appropriate numerical coefficients John S Werner and B R Wooten[36] propose a somewhat different nonlinear model, under which $(p\mathrm{M} + q\mathrm{L})^{3\ 5} - r\mathrm{S}$ yields the yellow-blue channel (where *p*, *q*, and *r* are the relevant coefficients) Either way, there is no threat to color objectivism So long as there is, for each channel, a well-defined function that connects the values of S, M, and L with the resultant color experience, that function, *linear or not*, will have an objectivist counterpart function connecting the relevant color with S*, M*, and L* (under normal viewing conditions) Thus, as before, given how the human visual receptors collectively operate, for each color experience, there is a complex surface disposition normally tracked by the experience—a complex disposition with which the color represented by the experience may be identified

---

[34] Letting x and y be variables over a domain, a function f is linear if and only if, for any real numbered constants, A and B, $f(Ax + By) = Af(x) + Bf(y)$ A function is nonlinear if and only if it is not linear Thus, a function involving exponents is nonlinear

[35] "A Model of Colour Vision for Predicting Colour Appearance," *Color Research and Application*, VII (1982) 95-112

[36] "Opponent Chromatic Mechanisms Relation to Photopigments and Hue Naming," *Journal of the Optical Society of America*, LXIX (1979) 422-34

We turn now to the unitary/binary distinction We shall assume that the oversimplified opponent-processing model is correct[37] This is done purely to simplify the discussion The points we make below apply, mutatis mutandis, to the more realistic versions of the model Let us grant, then, that the experience of red (not pure red) occurs if and only if $(L - M) > 0$, and the experience of yellow (not pure yellow) occurs if and only if $((L + M) - S) > 0$ Let us also agree with Hardin that necessarily, when something looks orange, it looks reddish and it also looks yellowish Indeed, let us accept that looking orange is one and the same as looking reddish and also looking yellowish Let us further suppose that looking reddish is just looking red without looking pure red and likewise for the case of looking yellowish The quantitative model now predicts that the experience of orange occurs if and only if $(L - M) > 0$ and $((L + M) - S) > 0$

Our corresponding (vastly oversimplified) objectivist claim about surface orange is as follows a surface is orange if and only if it has a surface reflectance that disposes it, under *normal viewing conditions* (NVC), to reflect light such that, for the reflected light, $L^* >_s M^*$ and $(L^* + M^*) >_s S^*$ Since, on the proposed account, a surface is reddish (that is, red but not pure red, as we are assuming) if and only if it has a surface reflectance that disposes it, under NVC, to reflect light such that, for the reflected light, $L^* >_s M^*$, and a surface is yellowish if and only if it has a surface reflectance that disposes it, under NVC, to reflect light such that, for the reflected light, $(L^* + M^*) >_s S^*$, orange is, quite literally, a mixture of reddishness and yellowishness The first disposition is constituted by the last two Orange, thus, is a binary hue

Note that this line of response does *not* commit the physicalist to holding that red is purply orange Looking purple is looking reddish and looking bluish According to the model, the experience of blue (but not pure blue) occurs if and only if $(L + M) < S$ So, the experience of purple occurs if and only if $L > M$ and $(L + M) < S$ Accordingly, a surface is purple if and only if it is disposed under NVC to reflect light such that $L^* >_s M^*$ and $(L^* + M^*) <_s S^*$ Evidently, then, red is not a literal mixture of purple and orange Indeed, no color is For no reflectance is such that it allows the same region of a surface to reflect light that satisfies incompatible conditions (Pure) red, according to the proposed account, is not a literal mixture of any other colors at all It is thus a unitary color

---

[37] From Hardin, *Color for Philosophers*

This is our preferred response to the unitary/binary objection to physicalist realism about colors But it should also be pointed out that there is no difficulty for the physicalist even if we understand the unitary/binary distinction in the way Hardin suggests For Hardin, as already noted, orange is reddish yellow in that necessarily something, in looking orange, looks reddish and looks yellowish By contrast, yellow is not reddish green, since it is not necessary that when something looks yellow, it looks reddish or greenish For unique yellow, when something looks *that* shade, it does not look red or green at all

Orange, then, is a perceptual mix of red and yellow, but yellow is not a perceptual mix of red and green This, we claim, the common-sense physicalist with respect to color can happily accept For facts about perceptual mixture are best taken as facts about how colors are represented in color experience In experientially representing something as orange, we represent it as being reddish and also as being yellowish But there is a shade of yellow such that, in experientially representing something as having that shade (unique yellow), we do not represent it as being reddish and being greenish Nor do we represent it as being orange and lime Thus, the determinables red and yellow enter into the content of the experience of something's looking orange, whatever the shade of orange But in the case of yellow, there is a shade such that the only determinable that enters into the content of something's looking that shade of yellow is yellow This tells us that there is an important difference in how we represent orange and yellow in our color experiences, a difference that is naturally explicable in terms of opponent processing But it presents no threat to the view that the colors, orange and yellow, are objective, physical properties of the same general type

Admittedly, on the assumption that our color experiences are generally veridical, the reply just offered entails that orange things *are* reddish and also yellowish That, however, is unproblematic for the reasons given earlier, if colors are objective dispositions of the sort we have proposed So, there is nothing in the trichromatic unitary/binary distinction that should lead us to give up common-sense color realism

Again, the extension to the tetrachromatic case is straightforward Let us give one illustration Consider the case of the goldfish According to the earlier model, an experience of H^+ (not pure H^+) occurs if and only if $((UV + S) - L) > 0$ Likewise, an experience of I^+ (not pure I^+) occurs if and only if $(UV + (S - M) + L) > 0$ Hence,

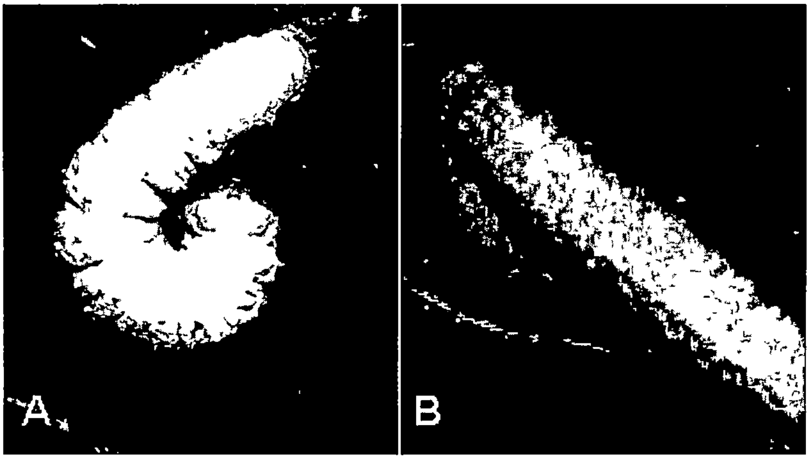


Figure 1 Caterpillar and Oak Leaf pictures images of the gray shoulder knot caterpillar (A) under UV illumination (300 - 400nm) and (B) in the human-visible spectrum (400 - 700nm)[38]

we claim, a surface has the hue, $H^+ + I^+$, if and only if it is disposed, under NVC, to reflect light such that, for the reflected light, the total quantity, $UV^* + S^*$, is significantly greater than L^* and the total quantity, $UV^* + S^* + L^*$, is significantly greater than M^* Since a surface is H^+-ish if and only if it is disposed, under NVC to reflect light such that the total quantity, $UV^* + S^*$, is significantly greater than L^*, the binary color, $H^+ + I^+$, is partly constituted by H^+-ishness in a quite literal way The ternary case is handled in a parallel manner Further, certain hues goldfish experience, for example, pure H^+, will be both objective and unitary

For the goldfish, then, the binary hues are conjunctive properties whose conjuncts are exactly two (nonconjunctive) hues accessible to the goldfish, for example, H^+-ishness and I^+-ishness The ternary hues are likewise conjunctive, but their conjuncts are three in number The unitary goldfish hues are those which are not conjunctive in the above way

IV THE CATERPILLAR AND THE LEAF

Consider again the caterpillar of the gray shoulder knot moth (lithophane ornithopus) Suppose that one such caterpillar is sitting on the underside of an oak leaf, as shown in figure 1 above

---

[38] Reproduced from Cuthill, Partridge, Bennett, Church, Hunt, and Hunt, p 186

To normal humans, the oak leaf and the caterpillar both look the same color in daylight (pure green, let us suppose[39]). But birds have the ability to discriminate between caterpillar and oak leaves based solely on color. Therefore, to normal birds, the oak leaf and the caterpillar do not look the same color in daylight. It is chauvinistic to say that we get the color right and birds get it wrong—after all, birds' characteristic interactions with caterpillars of this sort are beneficial to their viability. So, if we are to avoid arbitrarily giving preference to one species' color vision over another's, we are bound to say that the caterpillar and leaf are one color to humans, and two colors to the birds. How can this be?

There is no problem here for the projectivist. If colors are mental features wrongly projected onto the external world, then birds and humans project different colors onto the caterpillar, and both are mistaken in doing so. The caterpillar, in reality, is colorless. There is also no problem if colors are partly subjective, response-dependent dispositional properties. Since the typical subjective color response of a normal human perceiver to the caterpillar will be different from that of a normal bird to the caterpillar, the caterpillar will have both colors. But, how can the leaf and the caterpillar be objectively the same color and yet objectively different colors? A corresponding question arises in connection with the case of kestrel and human vision of straw covered with water versus straw covered with vole urine.

According to our (grossly oversimplified) account, an object is (pure) green if and only if it is disposed, under normal viewing conditions (NVC), to reflect light such that, for the reflected light, $M^* >_s L^*$ and $S^* =_a$ the total quantity, $M^* + L^*$. Since this condition is met (let us assume for the sake of simplicity) by the caterpillar and the leaf, both are pure green. Let us now hypothesize that the bird's color detection system is such that, viewing the caterpillar, the bird has an experience of a hue we may call 'UV$^+$-green' (hereafter UVG, for short). UVG has the following possession condition: an object possesses UVG if and only if it is disposed, under NVC, to reflect light such that, for the reflected light, $M^* >_s L^*$ and $S^* =_a$ the total quantity, $M^* + L^*$, and further the amount of reflected light in the UV portion of the spectrum (UV^*) is sufficiently large relative to S^*, M^*, and L^*. This is, of course, very vague with respect to UV^*, but for present purposes, we can leave the proposal in this incomplete form.

---

[39] This assumption is for simplicity only. In reality, very few objects outside the laboratory are pure green.

The caterpillar and the leaf, then, are both pure green—that is, they are green without being yellowish, bluish, or reddish—but the caterpillar also has the property UVG, whereas the leaf, given its failure to reflect much UV-light, lacks that color When we view the caterpillar and the leaf, our visual system picks up on the shared hue, pure green The bird's visual system is tetrachromatic, however, and it picks up on a difference in color that it is beyond our capacity to discern In a sense, then, colors are species specific But, as we emphasized earlier, this sense is perfectly compatible with their being objective

Perhaps it will be complained that it is extremely counterintuitive to suppose that the caterpillar is *both* pure green *and* UVG, that the caterpillar literally has *two* colors Our intuitions certainly tell us that nothing can be simultaneously red and green They also tell us that nothing can be simultaneously pure yellow and green In these cases, our intuitions are working within trichromatic color space, which is closed to novel hues The color, UVG, however, is not inside trichromatic color space—it is located in tetrachromatic color space We humans only have three types of cones in our eyes, and our lenses filter out UV light We simply lack the hardware to see in tetrachromatic color space Thus, when it comes to UVG, we have no clear idea what it is like to see this color, and we shall never have a clear idea of what it is like to see it Given this point, it is hard to see how our proposal is counterintuitive

Furthermore, it is not difficult to supply a possible conception of UVG that makes the claim we are advancing with respect to the color of the caterpillar intuitively very plausible The caterpillar surface, in being pure green, is green without being reddish or yellowish or bluish That is, it is green without having any other humanly accessible hue But, of course, that is compatible with supposing that it has a further humanly inaccessible hue such that the latter hue and green comprise UVG On this conception, UVG is a binary hue for the birds, one component of which is inaccessible to us Thus, just as a surface can be orange and reddish, so too a surface can be UVG and pure green And just as a red-green human dichromat would experience only the yellow in an orange surface and would have no clear idea of what it is like to experience orange, so too we trichromats experience only the green not tinged with any blue, red, or yellow, as we view the caterpillar, and we have no clear idea of what it is like to experience UVG[40]

PETER BRADLEY and MICHAEL TYE

Temple University

---

[40] For more on color realism, see Tye, *Consciousness, Color, and Content*, chapter 7

REVIEW ESSAYS

RECENT WORK IN PHILOSOPHY OF MATHEMATICS*

Naturalism in Mathematics PENELOPE MADDY New York Oxford University Press, 1997 viii + 254 p Cloth $55 00, paper $24 95 *Philosophy of Mathematics Structure and Ontology* STEWART SHAPIRO New York Oxford University Press, 1997 x + 279 p Paper $18 95 *Mathematics as a Science of Patterns* MICHAEL RESNIK New York Oxford University Press, 1997 x + 285 p Cloth $45 00, paper $19 95

The works under review are among the best written in recent years on the philosophy of mathematics Naturally, there will be differences among such diverse writers But even so, they display unifying themes These areas of overlap provide a stop-action photograph of the evolution of tacit attitudes among philosophers of mathematics, reflecting the dialectical pressures that common working assumptions are facing To focus this review essay, I shall follow out just one theme that shows itself in different ways in each the need for a detailed analysis of actual mathematical practice for the assessment of indispensability arguments

I INDISPENSABILITY ARGUMENTS AND FIELD
THE COMMON BACKGROUND

The basic outline of the indispensability argument is familiar we must regard mathematical claims as true because they are applied (apparently ineliminably) to learn truths in the natural sciences The history of this argument has contained few refined accounts of applications, though the argument seems to generate natural questions immediately Are these uses of mathematics necessary just because of physiological and psychological limitations of human reasoners? If yes, then why should this "narrowly practical" indispensability be taken as a guide to truth? If not, then just what kinds of advantage does mathematical reasoning contribute to broader scientific reasoning?

Of course, such discussions can be found in the literature, though they are rarer than one might hope It will be helpful to revisit the best-known example, replayed in the Stewart Shapiro volume (219-28) Hartry Field's effort to blunt indispensability arguments by prov-

---

* I am indebted to conversations with and comments from Andreas Blass, Igor Dolgachev, William Fulton, Robert Lazarsfeld, Penelope Maddy, Colin McClarty, Adam Morton, Mihnea Popa, Michael Resnik, and Larry Sklar

0022-362X/01/9809/488-97

ing a general result to the effect that (formal systems capturing) applied mathematical reasoning are conservative over "nominalistically acceptable" counterparts.[1] I shall approach this familiar topic with a different tack. A conservation result of this type that can be proven for an active domain of research will give us a foothold on how such results are viewed in practice. Gaisi Takeuti[2] showed some years ago that (properly formulated) analytic number-theory—exploiting the entire complex field—is conservative over elementary number-theory. This is interesting and important, but it did not prompt so much as a hint among practicing number theorists that analytic methods had been rendered obsolete. It is important to see why the extended theory is of value. The analytic theory allows us to make classifications the elementary theory does not. In particular, the Riemann zeta function can be defined in general and the Riemann hypothesis can be formulated.[3] This increased expressive power is one reason why proofs in analytic number-theory—like the first proof of the prime-number theorem on the distribution of primes[4]—are seen as "simpler."[5] This in turn supports the number theorist's sense that the classifications used in the extended theory are the genuinely *central* ones. Why should we draw any ontological or other philosophical consequences from the fact that the same theorems about numbers are provable in the restricted theory and the extended one, if the “dispensability” of the use of the complex numbers is of *only* this thin “in principle” sort?

---

[1] Another work by Shapiro—his recent introduction *Thinking about Mathematics* (New York: Oxford, 2000)—contains a useful summary of some of the issues, as does Field's collection *Realism, Mathematics, and Modality* (Cambridge: Blackwell, 1989).

[2] *Two Applications of Logic to Mathematics* (Princeton University Press, 1977). (In this case, neither the base theory nor the extended theory are "nominalistically acceptable" in Field's sense, but this will not matter for the points taken up here.)

[3] The most familiar restricted case of the zeta function is given by a series $\xi(s) = \Sigma_{i=1}^{\infty} 1/n^s$, which converges for complex numbers with real part greater than 1. G. F. B. Riemann showed how to extend the definition to almost the entire complex plane. The Riemann hypothesis is the (as yet undecided) conjecture that any z such that $\xi(z) = 0$ is such that the real part of z is ½.

[4] The "first proof" was provided by Jacques Hadamard and Charles de la Vallée-Poussin, subsequent proofs by Paul Erdös and Atle Selberg dispensed with the zeta function, these are generally described as "elementary" proofs, though in a different sense of 'elementary' from that sense relevant to Takeuti's conservativeness result.

[5] Note also that the simplification of the theory of the distribution of the prime numbers is gained by an *increase* in ontology. The complex numbers are uncountable, the integers are not. The availability of such conservation results has *not* prompted an abandonment of analytic number-theory on grounds of ontological parsimony. Far from it. Elementary number-theory, though in some sense simpler than analytic number-theory, is not simpler in at least some of the senses that in practice guide mathematical research. Nothing much hangs on this, but it is worth mentioning as a check against an overhasty assimilation of "simpler ontology" and "smaller ontology" which seems to be taken for granted in many philosophical discussions.

A variation on this point arises in connection to Field’s requirement that his nominalistic reformulations be “reasonably attractive ”[6] This constraint conjures up a cluster of questions that these two books engage in different ways [7] Where do our standards of attractiveness come from? (A priori philosophy, scrutiny of successful mathematical and scientific practice, or some weighted mixture?)

Of course, it is clear on its face that finer descriptive detail about mathematical practice is required here because it is beyond doubt that *some* techniques *are* practically indispensable just because of accidental quirks of human cognition In some obvious cases, the advantages of one framework over another do not seem *conceptually* deep or interesting, however needed they may be in practice, and interesting they might be to a cognitive psychologist [8] This points to a different way to frame a core question raised by the indispensability arguments Can *all* the advantages of mathematics be explained away as surmounting “mere medical impossibilities”? It seems to me unlikely, but a settled answer waits on a detailed analysis of a range of real cases [9] Just what does good mathematical work do for us?

---

[6] This extra codicil is crucial for the Field’s strategy to be interesting *Any* theory can be replaced by an equivalent “nominalistic” subtheory just *remove* (by brute force or Craig’s theorem) all sentences but those with the appropriately restricted vocabulary! But the resulting theories will be so madly disorganized as to be of no practical use No philosophical consequences would appear to follow from the observation that a good theory can in principle be replaced by a practically worthless theory, whatever philosophical charms that worthless theory might have

[7] Other questions arise, too, such as Why settle? Why not hold out for “maximally attractive”? Then conservativeness results will again be irrelevant, unless we appeal to a priori philosophical criteria of attractiveness rather than those which are tested in the scientific workshop

[8] For example, we prefer propositional systems with several connectives instead of the single Scheffer stroke, formulae in the one-connective system become impossibly cumbersome Some notational conventions make more practical sense than others (‘Let *M* be a module over the ring *R*’ rather than ‘Let *R* be a module over the ring *M*’) This collection of examples could obviously be expanded arbitrarily

[9] Elsewhere, Field appears to acknowledge that questions about the indispensability of mathematical reasoning will be more or less pressing depending on what kind of indispensability is at stake

> numerical theory is not only *harmless* for the purpose of drawing nominalistic consequences, it is also rather *useful* for this purpose There are other purposes for which this justification for feigning acceptance of numerical theory does not apply, and we must decide whether or not to genuinely accept the theory For instance, there may be observations that we want to formulate that we don’t see how to formulate without reference to numbers, or there may be explanations that we want to state that we can’t see how to state without reference to numbers *if such circumstances do arise, then we will have to genuinely accept numerical theory if we are not to reduce our ability to formulate our observations*

II RESNIK AND SHAPIRO STRUCTURE

The Shapiro and Michael Resnik books take a helpful first step in that both address the philosophical ramifications of the idea of *structure* (Resnik uses the word 'pattern') A plausible widespread opinion holds the emergence of the study of abstract structures to be distinctive of twentieth-century mathematics, to assess this opinion, we need to get clearer on just what "structure" is and to chart out the philosophical implications of the clarified concept [10]

Although the treatments in the two books have affinities, there is a difference in the role of indispensability argument Shapiro argues directly, as a motivation for his descriptive account of the role of structure in mathematics, that "the philosopher should not be content to note the apparent indispensability and then draw conclusions that spawn more questions than they answer By itself the indispensability argument has all the advantages of theft over honest toil" (245) By contrast, Resnik defends a "pragmatic" indispensability argument that takes an undifferentiated appeal to success as the basis of his epistemological account But Resnik too grants that the explanation of the success of mathematics must be part of a complete story, and that this will require an account of what counts as

> evidence that the theory of [posited entities] is consistent and bears fruitful connections to other mathematical and physical theories [since] successful positing is measured in terms of the success of some theory of the posits, where success is ordinarily measured by criteria that are local to the discipline What we want to know, then, is why mathematicians have been so successful in designing theories that count as successful by the criteria which the discipline of mathematics recognizes (195-96) [11]

---

or our explanations—"Platonism for Cheap?" in *Realism, Mathematics, and Modality*, pp 161-62, italics added)

It seems right that the potential need for mathematics in formulating observations or explanations involves different and more difficult issues than just abbreviating deductions Does mathematics in fact serve this function in science, and if so, how? At this point, the discussion will run out of gas without a more detailed descriptive and historical account of actual mathematics

[10] Shapiro defines a "system" as a collection of objects with certain relations (73) and a "structure" is "the abstract form of a system, highlighting the interrelationships among the objects, and ignoring any features of them that do not affect how they relate to other objects in the system" (74) As a definition, this would have shortcomings, but Shapiro puts it forward not as a definition strictly speaking, but rather as a kind of preliminary orienting device, to facilitate the "dialectical" elucidation of the notion of structure which is meant to unfold in the course of the book

[11] The practical situation facing the philosopher of mathematics is illustrated indirectly in Resnik's volume, via an argument that aims to undercut the

The relation of the Shapiro/Resnik picture to mathematicians' practice remains to be clarified, however This is true even with one of the most basic motivations for the "structuralist" view the slogan (repeated by Shapiro (241, and elsewhere)) that in practice, "a mathematician's interest stops at the level of structure up to isomorphism " This slogan has a grain of truth, and it captures many cases Certainly, practicing mathematicians do not trouble to ascertain whether the natural numbers are "really" the von Neumann or Zermelo ordinals [12] But there are also important cases where uniqueness up to isomorphism does *not* serve as a resting place even after a structure has been uniquely characterized, further efforts can be made to find the "right" definitions, "best" formulations, "most fruitful" concepts, and so on to facilitate particular applications For example, it was known throughout the twentieth century that the subject matter of complex function theory could be characterized in two equivalent ways via the Weierstrass definition of "analytic function" in terms of complex power series, or alternatively via the Cauchy-Riemann conditions Notwithstanding this sameness (up to isomorphism) of subject matter, the approaches to complex function theory associated with one or the other standpoint were, and remain, strik-

---

perception of a sharp distinction between mathematical entities and those of theoretical physics (chapter 6, 101-11) This interesting, substantive discussion, stands of necessity on the shoulders of foundational philosophy of physics To make the case about the entities of quantum field theory, Resnik draws on a range of superb philosophical discussions of the underlying physical issues There is no comparable reservoir of philosophical writing that is comparably informed by work on any area of contemporary mathematics outside of set theory

[12] For one striking example, note how the suggestion that it might matter what the real numbers are "made of" has been brusquely dismissed in a widely used introductory specialist's textbook In a chapter devoted to foundations, after proving the categoricity of the (second order) real closed field axioms, Michael Spivak writes

> This theorem brings to an end our investigation of the real numbers, and resolves any doubts about them There *is* a complete ordered field, and, up to isomorphism, only one complete ordered field It is an important part of a mathematical education to follow a construction of the real numbers in detail, but it is not necessary to refer ever again to this particular construction It is utterly irrelevant that a real number happens to be a collection of rational numbers, and such a fact should never enter the proof of any important theorem Reasonable proofs should use only the fact that the real numbers are a complete ordered field, because this property of the real numbers characterizes them up to isomorphism, and any significant property of the real numbers will be true for all isomorphic fields To be candid I should admit that this last assertion is just a prejudice of the author, but it is one shared by almost all other mathematicians—*Calculus* (Menlo Park Benjamin Cummings), pp 511-12

ingly different Even in the case of the reals, which is especially simple because of the existence of a *canonical* isomorphism, Michael Spivak's opinion is overstated in this respect the Dedekind cut and Cauchy sequence constructions exemplify two different properties completeness for lattices and completeness for metric spaces, respectively It is interesting and vitally important that these properties coincide for the specific structure of the real numbers, but they generalize in different directions

One paradigm of interest extending beyond isomorphism appears in Emil Artin's classic *Geometric Algebra*[13] Here, the isomorphism at issue is that between linear transformations of a space and the algebraic operations of matrices The anticomputational preference for the first is the basic motive for Artin's treatment as a whole, as he emphasizes in remarks of equal and opposite intemperance to the Spivak quote of footnote 12

> Mathematical education is still suffering from the enthusiasms which the discovery of this isomorphism has aroused The result has been that geometry was eliminated and replaced by computations Instead of intuitive maps of a space preserving addition and multiplication by scalars (these maps have an immediate geometric meaning) matrices have been introduced [This has produced] innumerable absurdities—from a pedagogical point of view (*ibid*, p 13)

It makes an important practical difference here which isomorphic representative is chosen Among the many reasons for this attitude, one in particular is worth highlighting it is (for many people, much of the time) easier to work with the "conceptual" characterization in terms of linear maps rather than with the "computational" framework of matrices For specific problems, with a well-chosen set of coordinates, a matrix representation may be useful, but the best choice of coordinates will differ from problem to problem Of course, it might be pointed out that this extra structure is interesting to mathematicians because it can be put to use to address further mathematical questions There is no such interest in the set-theoretic structure of one or another representations of the natural numbers because *that* extra structure is idle Fair enough—but in this form the observation becomes weakened to this near tautology mathematicians are interested in only those extra-isomorphic features which are mathematically interesting

This sort of case is especially interesting here because it leads naturally to some of the examples that are adduced as characteristic triumphs of the twentieth-century Bourbakist "structural" approach

---

[13] New York Interscience, 1957

André Weil[14] announced a crucial advance in algebraic geometry—his more "conceptual" approach to intersection multiplicity, replacing the previous, explicit algorithms from elimination theory—with a notorious open provocation "The device which follows , it may be hoped, finally eliminates from algebraic geometry the last traces of elimination-theory" (*ibid*, p 31) Weil's emphasis exemplifies, on a grander scale, the "anticomputational" stance displayed by Artin above Furthermore, an observer of the development of algebraic geometry in the middle third of the twentieth century could reasonably conclude that the ideology behind Weil's innovation had emerged triumphant Although the specific details of Weil's approach were quickly superceded and now seem archaic, his broader "structuralist" viewpoint seems to be vindicated in Alexandre Grothendieck's theory of schemes [15] The issue has become clouded in recent decades, with the explosion of computer resources prompting a renewed interest in explicit algorithms and an ability to exploit techniques that had not previously been feasible This has generated an energetic revival of a form of elimination theory (so-called Grobner bases) against the background of which Weil's tossed gauntlet could seem reactionary rather than revolutionary

Even a superficial one-paragraph history like the above reveals a cluster of tightly interwoven descriptive, methodological, and philosophical questions What is mathematically gained by the emphasis on "structure"? The particular success story of scheme theory also generates specific questions How typical is the motive of preserving dualities that scheme theory exemplifies (in the simplest cases)? What are its advantages? What differences between algebraic and geometric frameworks remain when we have taken into account the isomorphisms between corresponding structures? Since an acknowledged virtue of the theory of schemes has been the unification of algebraic geometry and number theory that it has supported, is there an interesting affinity with the sorts of unification of value in the natural sciences (Maxwell's electromagnetism, for example) that have been studied in general philosophy of science? Just to cite a few

Historical questions arise, too How has it come to be that a characteristic of twentieth-century mathematics has been the study of structure? What were the different pieces and how did they fall into place? Shapiro takes first steps in the direction of an answer, with a survey of some secondary literature on a few relevant nineteenth-century developments (chapter 5, 143-77) The chapter is useful for many reasons, not least because the material it discusses deserves to be better known There is more in the rich nineteenth-century developments that needs attention, however, before the emergence of the “structural” approach will have been made clear [16]

---

[14] *Foundations of Algebraic Geometry* (New York AMS, 1946)

[15] A clear and relatively accessible introductory survey of these developments is in Joe Harris, "Developments in Algebraic Geometry," *AMS Centennial Symposium* (Washington AMS, 1988), pp 89-100

By clarifying a philosophical concept of structure and adumbrating its metaphysical and epistemological outlines, Shapiro and Resnik have done a superb job of Lockean underlabor They say little about recent mathematics, however, so the (arduous) task of relating the philosophical investigation to the ideas of structure that are displayed in contemporary practice remains unfinished For a first step from the other side—a philosophically sensitive, detailed study of one aspect of ongoing mathematical research—I turn to Maddy's book

III MADDY, SET THEORY, AND “MATHEMATICAL NATURALISM”

I have been suggesting that the “implicit logic” of Shapiro's and Resnik's work pushes the philosopher of mathematics toward a fuller engagement with mathematical practice In this respect, Maddy's work represents an important advance, as its heart is a rich and illuminating investigation of one facet of mathematical activity the choice of candidate axioms in higher set theory It stands as a paradigm of philosophically thoughtful scrutiny of a genre of mathematics

Most of the critical response to Maddy's volume has been directed at the philosophical position—“mathematical naturalism”—that informs and motivates her “field anthropology” of set theory The mathematical naturalist holds that the standards of acceptable and unacceptable, justified or unjustified, “reasonably attractive” or unreasonably unattractive, and so on that should inform our assessment of mathematical theories are those standards internal to mathematical practice itself rather than those arising from philosophy or elsewhere For many “internal” questions—Is analytic number-theory simpler and more attractive than an elementary reformulation? Has scheme theory been successful?—such an appeal to the standards guiding mathematical practice appears plausible But for the issues that have been the stock in trade of recent philosophy of mathematics—Are the axioms of set theory *true?* Do we have reason to *believe* them?—this position faces pressures that are obvious on even cursory reflection With unanimity rare among philosophers, reviewers have scrambled to press a point that could be put as follows "Any number of complex organized practices may be successful by the standards constituting the practice itself—scientology, palm reading, I ching, Lurianic Kabbalah, and the bidding system of the Italian blue team, not to mention scores of conflicting systems of dogmatic theology We need a reason to regard mathematics as *more* than just one such practice among many Appealing to what mathematical standards say will not help us here "[17]

---

[16] Among the topics that recommend themselves are the different incarnations of the idea of “invariant” in algebra and Felix Klein's Erlangen programme, the related development of Sophus Lie's theory of continuous groups and their invariants, the twisted path of Otto Hesse's observation that 1-1 mappings of a domain induce “transfer principles”, Richard D Dedekind and Heinrich Weber's “structural” reinterpretation of Riemann's theory of algebraic functions and the subsequent career of Dedekind's insights in the hands of Emmy Noether and B C van der Waerden, Dedekind's epochal theory of ideals in his supplements to Peter Dirichlet's number theory textbook, Riemann's introduction of the concept of Riemann surface and Hermann Weyl's rigorous definition of the concept, and much more (The point of this list is not to accuse Shapiro of this or that omission, for the purposes of his book, the chapter he wrote was fine I aim only to give a hint of the richness of the soil that remains to be tilled)

Maddy is, of course, sensitive to this problem, and essays an apparently familiar answer "Mathematical naturalism" differs from a corresponding "astrological naturalism" because "mathematics is staggeringly useful, seemingly indispensable, to the practice of the natural sciences, while astrology is not As a result, one part of understanding science, within science, is understanding what mathematics is, what it does for science when it is used in application, and why it does this job so well" (205) [18] This answer has to be finely calibrated, since earlier Maddy rejected the indispensability argument as support for the truth of mathematical claims [19] Some commentators have suggested that this makes the subsequent appeal to applications untenable This complaint misses the deeper contours of Maddy's position Two points are combined (a) the methodological autonomy of at least some branches of mathematics vitiates the most straightforward indispensability arguments (in which, say, mathematics is confirmed as part of a hypothetico-deductive framework), but (b) the central role of mathematics in our scientific world view demands an account of the distinctive value of mathematics in any complete epistemology

---

[17] Variations on this theme have been urged in reviews by, among others, Bob Hale (*Journal of Symbolic Logic*, LXIV, 1 (March 1999) 394-96), Jill Dieterle ("Mathematical, Astrological, and Theological Naturalism," *Philosophia Mathematica*, VII, 2 (May 1999) 129-35), Gideon Rosen (*British Journal for the Philosophy of Science*, L, 3 (September 1999) 467-74), and Neil Tennant ("What Is Naturalism in Mathematics, Really?" *Philosophia Mathematica*, VIII, 3 (September 2000) 316-38)

[18] The appeal to applications comes after one further refinement Obviously, there is a deadly problem for astrology or any of these other systems if they issue in causal claims and predictions that conflict with the judgments of scientifically informed common sense The relevant comparison is therefore with a hypothetical "pure astrology" which makes no predictions about events in space and time

[19] Her rejection is prompted—characteristically—by descending from philosophical generalities to specific historical and scientific details In this light, she concludes the procedures for judging that reference to a mathematical entity is acceptable in pure or applied contexts differs crucially from those for admitting theoretical entities in physics and chemistry (like, paradigmatically, atoms) Furthermore, in scientific practice (as illustrated by the lag between the acceptance of atoms as theoretical conveniences and the acknowledgment of their existence) settling existence questions typically requires more than a quick game of spot-the-quantifier She reasonably takes this to defang the simple indispensability argument

Another familiar point appears as the dialectic cycles through again applications, it might be objected, give epistemological credentials only to the *applicable bits* of mathematics, and not (say) large cardinal axioms Maddy responds thus "my mathematical naturalist sees mathematics as a unified undertaking which we have reason to study as it is" (205 footnote 15) The discussion obviously cannot rest here One thing investigations of mathematical practice should address is whether or not mathematics *really is* such a "unified undertaking " What *is* the interaction among the wide variety of styles of reasoning and subjects which are—as a matter of institutional fact—clustered together in mathematics departments? Maddy's book emphasizes one—possibly idiosyncratic—domain the program of the (self-titled) "Cabal seminars " It is not inconceivable that for directly applied areas of mathematics, this work might be as hermetic as the name implies This point, too, has been pressed by Maddy's critics, but it does not seem to be appreciated that the obvious reply preserves and extends the spirit of her work (though this reply is independent of mathematical naturalism as a philosophical view) Of course, it *could* be that mathematical practice is *not* tightly interconnected Perhaps a detailed investigation will reveal it to fragment into discrete solitudes But this can be ascertained only with *more* work, not less, of the sort Maddy has given us Further studies of the sort Maddy has provided—for other flavors of set theory as well as algebra, geometry, number theory, "hard" analysis, and so on, plus their intramural relationships and applications in physics and elsewhere—are here a *sine qua non* In brief, we need to learn more about mathematics to make progress on this front in the philosophy of mathematics

JAMIE TAPPENDEN

University of Michigan/Ann Arbor

NOTES AND NEWS

The editors regret the untimely death of Wesley C. Salmon, University Professor Emeritus of Philosophy and Professor Emeritus of the History and Philosophy of Science at the University of Pittsburgh, on April 22, 2001. Born in 1925, Salmon earned his Ph.D. under Hans Reichenbach in 1950 at the University of California/Los Angeles. He held appointments at several universities before serving as Norwood Russell Hanson Professor in the Department of History and Philosophy of Science at Indiana University/Bloomington, 1963-1973. He worked at the University of Arizona/Tucson, 1973-1981, and then joined the University of Pittsburgh faculty in 1981, where he remained until his retirement in 1999. He served as resident fellow in the Center for Philosophy of Science, and from 1983, he held the rank of University Professor of Philosophy. Salmon's books and articles have ranged broadly over the theory of scientific explanation, causality, probability, scientific confirmation and induction, and the philosophy of physical science. His well-known books include *The Foundations of Scientific Inference* (1967), *Scientific Explanation and the Causal Structure of the World* (1984), *Four Decades of Scientific Explanation* (1990), and *Causality and Explanation* (1998)—a collection of essays spanning several decades. Salmon's career was distinguished by an array of academic recognitions, including the presidencies of the American Philosophical Association (Pacific Division, 1977-78) and the Philosophy of Science Association (1971-72).

The editors regret the death of Alice Ambrose Lazerowitz, Professor Emeritus of Smith College, on January 25, 2001, at the age of 94. Born in 1906, Lazerowitz received her first Ph.D. from the University of Wisconsin and another from Cambridge University, where she studied under G. E. Moore and Ludwig Wittgenstein. In addition, she earned her LL.D. from Millikin University in 1958. Lazerowitz began her career at Smith in 1937, remaining at the college and becoming Professor Emeritus in 1972. She was awarded the Austin and Sophia Smith chair in Philosophy in 1964. Her publications include *A Controversy in the Logic of Mathematics* (1933), *Essays in Analysis* (1966), and *The Changing Face of Philosophy* (1968). She collaborated with her husband, Professor Morris Lazerowitz, on a number of works: *Fundamentals of Symbolic Logic* (1948), *Logic: The Theory of Formal Inference* (1961), *Philosophical Theories* (1976), *Essays in the Unknown Wittgenstein* (1984). In more recent years, Professor Lazerowitz edited the publication of her notes (and Margaret Macdonald's) which were based on lectures given by Moore and Wittgenstein at Cambridge. Lazerowitz served as President of the Eastern Division of the American Philosophical Association (1975–76) and as an editor for the *Journal of Symbolic Logic*.

THE JOURNAL OF PHILOSOPHY

VOLUME XCVIII, NO 10, OCTOBER 2001

05/12/07

COINING TERMS IN THE LANGUAGE OF THOUGHT
INNATENESS, EMERGENCE, AND THE LOT OF CUMMINS'S ARGUMENT AGAINST THE CAUSAL THEORY OF MENTAL CONTENT*

Robert Cummins[1] argues that any *causal theory of mental content* (CT) founders on an established fact of human psychology that theory mediates sensory detection He concludes "CT is false in all its forms" (*ibid*, p 541) Cummins takes his conclusion to follow from general characteristics of human sensory systems and CT, thus, he does not discuss in detail the nature of the theories that mediate sensory detection or the nature of the mental representations that, according to CT, are tokened as the result of humans' sensory interactions with the world I contend that the details matter when we fully appreciate the range of ways in which mental content and the *vehicles* of mental content may plausibly emerge in the human cognitive system, CT appears, after all, to be consistent with theory's mediating role in human perception

---

* I would like to thank Edward Averill, Robert Cummins, and J Robert Loftis for helpful comments on an earlier draft of the present paper I also thank, for their constructive feedback, audiences at Texas Tech University and the Northwest Conference on Philosophy (1999), to whom I presented earlier versions of this work I would like to acknowledge particularly the contributions of Frederick Suppe, Aaron Meskin, J Robert Loftis, Mark Webb, and Dan Kaufman (Texas Tech) and Todd Jones, David Beisecker, Hollibert Phillips, David DeMoss, and Camille Coffey (Northwest Conference) In my early thinking about nativism, I benefited greatly from conversations with Charles Chastain, especially our discussions of Fodor's nativist views

[1] "The LOT of the Causal Theory of Mental Content," this JOURNAL, XCIV, 10 (October 1997) 535-42

0022-362X/01/9810/499-530

I CUMMINS'S ARGUMENT AGAINST CT

Generally speaking, CT asserts that the content of a primitive mental representation, or term in a *language of thought* (LOT),[2] is fixed by its causal relation to kinds or their individual members (or properties or their instantiations—assume this qualification throughout) Proponents of CT[3] have formulated a variety of complex theories of content, each theory specifying the content-fixing causal relation in its own way Given the variety of such approaches, it is not entirely clear how properly to circumscribe causal theories, or to what extent there is a substantive philosophical division waiting to be made between true causal theories and the rest This is largely a side issue, however Cummins's critique takes aim at one specific feature shared by many theories of content, a feature he places at center stage when describing CT He defines CT as "the doctrine that the contents of the semantic primitives in the human scheme of mental representation are determined by their role in detection" (*op cit*, p 535) where, in particular, the determination of content rests on the subject's ability to detect *reliably* members of the kind in question Thus, whether we classify a given theory of content as a causal theory matters little for present purposes Cummins's critique of CT simply takes as its target any theory for which reliable detection is a necessary condition of correct content assignments to human mental representations, the literature contains enough theories of this sort to merit careful and critical consideration of Cummins's argument

Cummins claims that the mediating role played by theory in human perception creates an insuperable problem for CT Currently accepted theories of human perception imply that, for a subject reliably to detect members of a kind, she must possess a theory of that kind (note well this attribution of theory does not imply that the subject consciously holds beliefs with the content expressed by the theory) Cummins labels this condition on detection the *nontransduc-*

---

[2] See Jerry Fodor, *The Language of Thought* (Cambridge Harvard, 1975) Sometimes, when it seems convenient for expository purposes, I refer to mental representations as "concepts," without thereby intending to take any position regarding the nature or existence of concepts as abstract, as opposed to psychological, entities

Whether LOT exists is a contentious issue Cummins argues that CT entails at least a minimal version of the LOT hypothesis (*op cit*, pp 539-41) While his argument merits attention, I set it aside here I do not defend CT by attempting to sever it from LOT, as causal theorists might propose to do in response to Cummins's critique of CT, consequently, for present purposes, I grant LOT's existence

[3] Fred Dretske and Fodor have developed two of the most prominent causal theories, see Dretske's *Knowledge and the Flow of Information* (Cambridge MIT, 1981), and *Explaining Behavior* (Cambridge MIT, 1987), and Fodor's *Psychosemantics* (Cambridge MIT, 1987), chapter 4, and *A Theory of Content and Other Essays* (Cambridge MIT, 1991), chapter 4

ibility of distal properties (NTDP) A difficulty arises for CT because, in many cases, the theory required to mediate reliable detection must be explicitly represented and learned In such cases, CT cannot explain how the subject comes into a content-fixing causal relation to the relevant kinds or their members to construct a theory that will allow her reliably to detect *C*s, the subject must possess a concept with the content *C*, so that *C* can appear in the subject's explicitly represented theory of *C*s Yet the subject cannot possess the concept *C* in the absence of a theory that allows her reliably to detect *C*s Without the theory, there is no concept, without the concept, the subject cannot formulate the theory Thus, CT fails as a general theory of content for human mental representations Cummins conveys this point by way of an example the typical human can reliably detect cats in the environment but does not do so directly, much cognitive processing comes between sensory stimulation and the subject's tokening of a LOT term with the content *cat* As Cummins puts it "To detect cats requires a theory that says, in effect, what sorts of proximal stimuli are reliable indicators of catness To detect cats visually, you have to know how cats look" (*op cit*, p 536) Alas, the subject can have no explicit knowledge of how cats look *sans* the ability to represent cats Thus, for many human concepts of just the sort to which CT is supposed to apply, CT does not provide a viable theory of content, together with the facts of human perceptual psychology, CT implies that you must possess a concept of cats as a means of acquiring a concept of cats, which is impossible

II INNATE AND UNLEARNED LOT TERMS

Cummins restricts the scope of the preceding argument if there are LOT terms whose contents are fixed independently of explicit-theory acquisition, perhaps a causal theory of content applies to such terms Here Cummins acknowledges the possibility that some LOT terms are innate, their content fixed independently of their appearance in explicit theories (Cummins suggests 'square' as a candidate—*op cit*, p 537) Still, causal theorists should postpone the celebration, given what would appear to be the limited stock of innate terms, Cummins considers it unlikely that nativism will rescue CT (*op cit*, p 537)

It is intuitively plausible that humans possess a relatively limited range of innate concepts, yet we should wonder whether this intuitive reaction can yield the result Cummins claims for it It is notoriously difficult to pin down the precise meaning of nativist claims,[4] which of course makes it difficult to evaluate such claims What is more, the issue of innateness bears only indirectly on the present debate When evaluating Cummins's argument against CT, the category of importance to the causal theorist—the category of terms that escape Cummins's attack—consists of terms that acquire content in the absence of explicit theories of those things to which the terms apply The force of Cummins's critique of CT rests squarely on cases where explicit theory mediates detection, where explicit theory must be present if the content of the pertinent LOT terms is to be fixed per CT The greater the number of primitive LOT terms that can be acquired in the absence of explicit theory, the less destructive the causal theorist should find Cummins's argument Let us use *learned terms* to refer to those LOT terms for which, assuming CT and NTDP, it is a necessary condition of their acquiring content that the human subject formulate an explicit theory of that to which the terms apply, call all other LOT terms *unlearned* The present point can then be expressed as follows we should not assume that the categories of unlearned terms and innate terms are coextensive[5] Although some definitions of "innateness" may be weak enough to apply to all unlearned terms, Cummins cannot rest happily defining "innateness" in a permissive way The specific purposes to which Cummins puts antinativist claims demand a strong definition of "innateness " When, for example, Cummins points out the obvious poverty of our innate conceptual endowment, he appeals to an intuition whose force rests on a strong notion of innateness it is because we have in mind a demanding notion of innateness that the poverty of innate concepts seems so clear If we limit our discussion to the more demanding definition of "innateness" Cummins's arguments require, we have no reason to think that innate terms exhaust the contents of the class of unlearned terms In what follows, I shall have a good deal to say about innateness and the role antinativist claims play in Cummins's critique of CT Nevertheless, we should not lose sight of the more fundamental point concerning the scope and composition of the category of unlearned terms

---

[4] For an extended discussion of this problem, see Fiona Cowie's *What's Within Nativism Reconsidered* (New York Oxford, 1999), part I, also see A Karmiloff-Smith,

---

K. Plunkett, M H Johnson, J Elman, and E A Bates, "What Does It Mean to Claim that Something Is 'Innate'? Response to Clark, Harris, Lightfoot and Samuels," *Mind and Language*, XIII (1998) 588-97

[5] Arthur C Danto makes a similar point in his "Semantical Vehicles, Understanding, and Innate Ideas," in Sidney Hook, ed , *Language and Philosophy* (New York University Press, 1969), pp 122-37 "Obviously, the move from *unlearned* to *innate* is abrupt, inasmuch as learning is but one avenue of acquisition " (p 128)

Sections III and IV establish a general conclusion the category of unlearned terms seems likely to extend much further than Cummins acknowledges These sections do not alone rebut Cummins's critique of CT, for no matter how large the category of unlearned terms, there likely remain at least some primitive, learned LOT terms, if CT is to possess a satisfying degree of generality, the causal theorist must explain how causal relations can fix the content of primitive, learned LOT terms while avoiding the apparent circularity Cummins identifies To this end, I argue in section V that Cummins underestimates the theoretical resources available to the causal theorist in her attempt to explain LOT term acquisition The initial framing of theoretical axioms that facilitate detection of *C*s should be understood to include only a *nonsemantically individuated* LOT term '*C*' Individuating a newly coined LOT term according to content-independent criteria allows us to see how a primitive LOT term can be introduced and its meaning fixed by theoretically mediated detection, without the term's having to possess content as a condition of its appearance in theoretical axioms

III INNATENESS AND THE EMERGENCE OF CONTENT

While Cummins accepts that some LOT terms may be innate, NTDP is non-negotiable, theory mediates even the causal relation between innate terms and their contents What is the nature of such theory? Are infants born possessing a theory of squares, sufficiently detailed and accurate to facilitate the reliable detection of squares? In a footnote, Cummins remarks on mediating theories from the standpoint of the LOT theorist "According to LOT, if it [the mediating theory] is learned, it will be a set of sentences in LOT If it is innate, it might, in some sense be implicit in the architecture" (*op cit*, p 536, footnote 5) The former case drives Cummins's negative argument, yet the latter deserves more attention For the causal theorist wishing to lessen the force of Cummins's critique, the idea of an implicit theory seems to hold promise If innate, implicit theory can, in the case of 'square', play the detection-mediating role required by CT and NTDP, why should not implicit theory, *innate or otherwise*, play such a role in the case of 'cat'?

Cummins briefly considers the idea that innate, implicit theory plays the requisite mediating role in the case of 'cat', but he cannot accept the implications of this view In his words "[W]e are not born knowing how cats look We must, then, learn the theory that mediates cat recognition" (*op cit*, p 537) To give the discussion focus, I offer the following reconstruction of Cummins's response to the nativist defense of CT

(P1) CT can only account for the fixation of the content of the primitive LOT term 'cat'[6] by appealing either (a) to what the human subject knows at birth about cats or (b) to the human subject's acquisition of an explicit theory of cats

(P2) What the human subject knows at birth does not fix the content of her primitive LOT term 'cat'

Therefore, an explicit theory of cats must be acquired, if CT provides the correct account of how the human subject acquires the primitive LOT term 'cat'

As Cummins presents the argument, and as I have formulated it, the conclusion pertains specifically to 'cat' Yet for the argument to fulfill its purported function, it must apply to a wide range of primitive LOT terms Cummins intends his overall argument to constitute a decisive critique of CT, a single counterexample to CT, while providing a challenge to causal theorists, would hardly devastate their research program Presumably, then, Cummins chooses 'cat' somewhat arbitrarily, 'cat' standing in for a large class of primitive LOT terms all of which, assuming CT and NTDP, can only acquire content via the mediating effects of explicitly represented theory (and thus, according to Cummins, are such that CT cannot account for the fixation of their content) In light of the generality of his point, one should at least begin to doubt that (P1) canvasses all of the ways in which a causal theorist might plausibly account for the fixation of the content of a primitive term in a human's LOT In the subsections that follow, I offer specific reasons to think that (P1) fails to cover all of the relevant possibilities, or that, if it does cover them all, it does so by presuming a weak interpretation of what it is to be innate or to know something at birth—so weak that (P1) and (P2) cannot legitimately collaborate in support of Cummins's conclusion

III A Cummins rightly allows that innate, implicit theory might play the mediating role required by the conjunction of CT and NTDP Why, one may wonder, should the content-fixing power of implicit theory be limited to *innate* implicit theory? If implicit theories that mediate reliable detection can appear over time in the postnatal subject, (P1) does not cover all cases Let me dub as *EI-terms* ('*E*' for emergent, '*I*' for implicit) all primitive LOT terms whose contents are fixed, per CT, via the mediating effects of implicit theory that is not present at birth That the human cognitive system contains EI-terms seems quite plausible, given the variety of ways in which information can be implicit[7] and the variety of ways in which the human cognitive and biological systems change over time At the very least, implicit theories could emerge as "maturational changes" themselves dependent on the organism's prior interaction with its environment (for that matter, an implicit theory could appear as the result of a subject's suffering a head injury—the effect of a certain kind of unpleasant interaction with the environment) [8]

---

[6] Following Cummins, I set aside the possibility that 'cat' is compound Cummins rightly sees it as very unlikely that other LOT primitives define 'cat' (*op cit*, p 537)

In fairness to Cummins, we should pay careful attention to his description of implicit theory it is "in some sense, implicit in the architecture," according to CT-cum-LOT [9] If we take cognitive, or functional, architecture to be fixed, and if by this we mean unchanging, then no implicit theory at all can emerge, emergence is a form of change Cummins may well have this conception of functional architecture in mind Elsewhere in his recent writings,[10] when Cummins talks of functional architecture, he refers to it as "fixed," citing the work of Zenon Pylyshyn.[11] Cummins there takes Pylyshyn's view to be representative of the computationalist's perspective, a perspective we can safely assume he attributes to the causal-cum-LOT theorist. The sense in which Pylyshyn wishes to hold fixed the functional architecture will not, however, do the duty Cummins expects of it.

---

[7] See Cummins's "Inexplicit Information," in Myles Brand and Robert M Harnish, eds , *The Representation of Knowledge and Belief* (Tucson Arizona UP, 1986), pp 116-26, especially pp 119-20 Of special interest here is Cummins's discussion of domain-implicit information, which seems especially likely to emerge over time in the subject Here one should also take into account much recent work on embedded cognition This work encourages a minimalist view of the amount of internal, explicit representation necessary for a cognitive system to exhibit a full range of beliefs or possess a standard complement of knowledge or information Those who view cognition as embedded claim that, in a surprising number of cases, we best explain human behavior and cognitive capacities without appealing to a rich set of internal representations and complex internal computational processes manipulating those representations Although some internal representations are admitted in these cases, it is claimed that the standard subject's cognitive states and capacities are, to a great extent, "off-loaded" (that is, stored in or heavily dependent upon the environment), and thus emerge from the relations the subject bears to her environment and its structure If implicit theory is embodied in the relations themselves, it is easy to see how implicit theory could emerge over time, as the subject *comes to bear* the relevant relations to her environment On embedded cognition, see Andy Clark, *Being There Putting Brain, Body, and World Together Again* (Cambridge MIT, 1997), Ron McClamrock, *Existential Cognition Computational Minds in the World* (Chicago University Press, 1995), David Houghton, "Mental Content and External Representations," *Philosophical Quarterly*, XLVII (1997) 159-77, Clark and David Chalmers, "The Extended Mind," *Analysis*, LVIII (1998) 7-19, Mark Rowlands, *The Body in Mind Understanding Cognitive Processes* (New York Cambridge, 1999) Some of this work argues for the stronger thesis that cognition is extended states and processes of the human cognitive system literally comprise parts of the world beyond the boundary of the organism While more tendentious than the claim that human cognition is embedded, the view that cognition is extended also supports the present point

[8] Cf Fodor, "The Present Status of the Innateness Controversy," in *Representations* (Cambridge MIT, 1981), pp 257-316, here p 275

[9] "The LOT," p 536

[10] *Representations, Targets, and Attitudes* (Cambridge MIT, 1996), p 31 Also note Cummins's citation of Pylyshyn's work when discussing functional architecture and transduction ("The LOT," p 536, footnote 4)

Pylyshyn holds functional architecture fixed for explanatory purposes:

> [T]he architecture must form a cognitive "fixed point" so that differences in cognitive phenomena can be explained by appeal to arrangements (sequences of expressions and basic operations) among the fixed set of operations and to the basic resources provided by the architecture. If the functional architecture were to change in ways requiring a cognitive rule-governed explanation, the architecture could no longer serve as the basis for explaining how changes in rules and representations produce changes in behavior (*ibid.*, p. 114; also see p. 131).

Appeals to functional architecture are supposed to yield explanatory power by *accounting for* cognitive processes. Such appeals are intended to explain semantically characterized processes in terms of series of applications of nonsemantically characterized, simple operations, which operations are not themselves affected by processes described in semantic or representational terms (using Pylyshyn's now well-known description, these operations are not "cognitively penetrable").[12] This provides an important sense in which the functional architecture is fixed: functional architecture is not subject to changes that must be characterized in cognitive terms; instead, functional architecture's being mere physical mechanism, following laws of nature stated in nonsemantic terms, explains how cognitively characterized behavior can arise from, or be instantiated in, an underlying physical system.[13]

The preceding picture of functional architecture and its role as an explanatory fixed point does not require that each aspect of the functional architecture be either innate or unchanging. Pylyshyn describes well the way in which the architecture can change and the way in which such changes can influence higher levels, that is, the symbolic (sometimes called "syntactic") and ultimately the semantic levels.

---

[11] *Computation and Cognition: Toward a Foundation for Cognitive Science* (Cambridge: MIT, 1984).

[12] Of course, this explanatory approach is not the only one possible. For some purposes, it may suffice to classify as basic certain semantically characterized processes and representations, explaining all other cognitive capacities by appeal to operations and arrangements of the basic, semantically characterized ones. This approach, however, does not allow an account of the workings of the entire cognitive system in noncognitive, nonsemantic terms.

[13] Among its other virtues, Pylyshyn takes this explanatory gambit to hold promise of at least a token reduction of cognitive processes (*op. cit.*, pp. 132-34).

> Thus biological factors interact with symbol level generalizations by causing modifications in the basic computational resources, which I call the *functional architecture*. Such modulations may be due to biochemical influences, maturation, the triggering effect of environmental releasers, dendritic arborization, atrophy of neural function from disuse, and so on (*op. cit.*, p. 259).[14]

Cummins's view seems insensitive to the possibility Pylyshyn describes: the possibility of a change in functional architecture that does not result from cognitive causes. Consider Cummins's comment on the representational stock and its relation to architecture:

> What Fodor has given us is an argument that primitive *representations* cannot be learned, that is, that one's basic representational power cannot be altered by learning. With this I am in complete agreement. One's representational repertoire is part of one's innate architecture.[15]

Cummins seems to assume that, if a concept is not acquired by cognitively described means (as the result of learning, in particular), then the concept is innate. We should reject this assumption, given that (a) Cummins takes a concept's innateness to entail knowledge, present at birth, about that to which the concept applies, (b) theory implicit in the architecture can mediate content fixation (that is, play the role of knowledge, in "knowledge, present at birth"), and (c) architecture can change postnatally in ways that do not consist in learning but nonetheless result in the subject's holding new implicit theories.

Pylyshyn's promising explanatory strategy places a certain constraint on our characterization of the functional architecture: in order that we satisfactorily explain cognitive processes by appeal to the elements of the cognitive architecture, we should not characterize that architecture itself in cognitive terms. Cummins's view might seem to cohere well with Pylyshyn's; we cannot explain the acquisition of new primitives *as the result of learning*, for this would undermine

---

[14] See also *op. cit.*, p. 260, where Pylyshyn identifies control structures as part of the functional architecture. In "Inexplicit Information," Cummins argues that information can implicitly reside in control structure (and, presumably, enough such information could constitute implicit theory). If functional architecture can change over time, and control structure is part of the functional architecture, and control structure can instantiate implicit theory, then implicit theory need not be innate in the sense of being present at birth.

[15] *Representations, Targets, and Attitudes*, p. 145.

Pylyshyn's explanatory strategy But Pylyshyn's view, and I would claim the computational approach in general, should allow the appearance of new "representational" primitives *as the result of noncognitive causes*

According to Pylyshyn, a representational primitive can be identified with "a physical code or a symbol" (*op cit*, p 26)[16] To count as part of the functional architecture, it must be individuated in the former terms "Being instantiated in the functional architecture merely means being explainable without appeal to principles and properties at the symbolic or the semantic level" (*op cit*, p 132) While it is useful to organize a discussion of the cognitive system around different levels—biological, syntactic, semantic—the present context demands investigation of interlevel relations In this regard, Pylyshyn argues that changes at the physical level alter the resources available to be described at the cognitive level, primitive operations and representations can emerge in the system largely as a result of changes at the physical level, but the physical changes themselves have no relevant cognitive or semantic characterization (I develop this idea in more detail below in sections IV and V) Cummins can, if he likes, deny the sort of interrelation Pylyshyn describes He might wish to claim that all implicit theory ever mediating the fixation of content in keeping with CT is fixed in the cognitive architecture at birth, never changing Given that the acceptability of (P1) depends on such a strong claim, however, Cummins had best offer significant considerations in its favor, he has not done so If, on the other hand, he countenances the sort of interrelation between levels that Pylyshyn describes (where new architectural features can emerge from matu-

---

[16] For two prominent computationalists' expression of a similar view, see Allen Newell and Herbert A Simon, "Computer Science as Empirical Inquiry Symbols and Search," in John Haugeland, ed, *Mind Design II Philosophy, Psychology, and Artificial Intelligence* (Cambridge MIT, 1997), pp 81-110, on p 85 (reprinted from the *Communication of the Association for Computing Machinery*, XIX (March 1976) 113-26)

In the passages cited here, Newell and Simon—and Pylyshyn—speak specifically of the nature of explicit representational primitives, and one might wonder about the connection between the postnatal emergence of such representations and our present concern changes in functional architecture that result in the emergence of new implicit theories The relation between the two issues is twofold First, note our position in the dialectic, at this point, I intend to establish the possibility that functional architecture can change postnatally To do so, it will suffice to have established any form of postnatal change in functional architecture, including as a possibility, the emergence of primitive explicit representations Second, we should also consider the possibility that, at least in some cases, it is at least partly via this very kind of change, that is, the emergence of primitive representations, that a new theory implicit in the architecture can come to be held by the subject ration, experience-dependent synaptic growth, and so on), (P1) fails to cover the full range of content-fixing options Thus, lacking some reason to think that all the changes in functional architecture amount to the unfolding of knowledge present at birth, human possession of EI-terms would seem to be a genuine possibility [17]

III B The preceding has shown that a theory implicit in the functional architecture need not be innate in the way Cummins's argument presupposes it to be We should also wonder why he limits implicit theory to aspects of cognitive architecture I do not suppose that he intends "theory implicit in the architecture" simply to mean *whatever* in the subject facilitates reliable detection in the absence of explicit theory But if he has in mind a substantive constraint on what counts as implicit theory built into the functional architecture, he should explicitly state the criterion and argue for it To the causal theorist, any such criterion will appear suspect To say that, relative to a given LOT term, an implicit theory facilitates content-fixing reliable detection, need, for the causal theorist, amount to no more than saying that, in the absence of an explicitly represented theory of that to which the term applies, the term nevertheless comes to stand in the content-fixing causal relation to the kind or to instances of the kind in question [18] The causal theorist might take great interest in the mechanisms that constitute implicit theory, the operation of which contributes to the fixation of the content of unlearned, primitive LOT terms, she might also be curious to what extent these mechanisms are embodied in the cognitive architecture. Regardless, she should be perfectly satisfied with my minimal characterization of implicit theory, for as regards the fixation of content, all that ultimately matters to the causal theorist is the establishment of the right causal relation. Given that CT seems to require no more than this liberal understanding of implicit theory, there seems little reason for the causal theorist to think that all content-fixing implicit theory is part of the cognitive architecture; at least Cummins has given her no reason to do so.

---

[17] One could claim that all such architectural changes do, in fact, amount to the unfolding of knowledge present at birth, however, in order to legitimate such a claim, one would have to offer a supporting view of what it is for knowledge to be present The following view—echoing one of the nativist targets of Book I of John Locke's *Essay concerning Human Understanding* (P H Nidditch, ed (New York Oxford, 1975), see chapter 2, §5)—would suffice knowledge of *As* is present in a subject if and only if either (a) the subject explicitly represents knowledge of *As*, (b) the subject is so constituted that, given certain experiences, she will someday acquire a concept with the content *A* that will appear in an explicit representation of knowledge about *As*, or (c) the subject is so constituted that, given certain experiences, she will someday acquire implicit knowledge of *As* This view justifies the claim that all primitive, unlearned LOT terms refer to things about which subjects know something at birth, notwithstanding postnatal changes in architecture, yet it does so at great cost We know something at birth about all things—including cats—some knowledge of which we will never explicitly or implicitly represent Cummins must reject this strategy, this cannot be the sort of knowing he has in mind when he treats the claim that 'We are not born knowing how cats look' as standing in need of no support He might have some other, general view in mind that would legitimate the claim that all physical changes in functional architecture amount to the unfolding of knowledge present at birth, but I can not make it out It is difficult to see how any view could be permissive enough to support the claim all architectural changes amount to the unfolding of knowledge present at birth, while being at the same time restrictive enough to support Cummins's antinativist views regarding our knowledge of that to which such concepts as 'cat' apply

[18] The causal theorist might, with some justification, approach these matters differently She might acknowledge substantive constraints on what is to count as implicit theory, but also argue that NTDP is false, that there are cases where we gain

III C. I have argued that implicit theory, construed as part of the functional architecture, need not be innate (section III A); I have also questioned Cummins's identification of implicit theory with what is built into the functional architecture (section III B). In doing so, I have tried to pry apart claims about innateness, implicit theory, and functional architecture. But we should not neglect what would seem to be Cummins's underlying reason for discussing theory implicit in the architecture. He assumes that any unlearned concept is innate and that a concept's innate status implies the presence of knowledge[19] at birth; the discussion of implicit theory, built into the

---

nothing by characterizing in semantic terms (as, say, bits of theory implicit in the functional architecture) whatever mediating machinery facilitates reliable detection. (This point was suggested by a comment of Aaron Meskin's.) I have, however, chosen to accept NTDP for the sake of argument and to criticize Cummins's position on other grounds.

[19] For an important qualification, see "The LOT," p. 536 (also, *Representations, Targets, and Attitudes*, p. 132). Here Cummins acknowledges his use of 'know' to be unorthodox in some respects, for it is the cognitive scientist's use, rather than the philosopher's. Whereas the philosopher's use implies the truth of what is known and also that it is justified, to know *P* in the cognitive scientist's sense, the subject need only treat *P* as if it were reliable information (that is, the subject need only believe *P*, in the not very demanding sense of 'believe' that allows you to believe that *P* without any conscious access to the content of *P* and without your being able to employ the belief that *P* in a wide range of contexts). For present purposes, Cummins's semantic clarification is moot. To be true to traditional philosophical usage, some may wish to substitute 'believe' (or given the lack of any requirement of conscious access, 'believe*', or, following Noam Chomsky, 'cognize') in place of 'know' when we encounter the latter in Cummins's argument. Then, however, I shall simply translate the argument in the text into talk about believing (or believing*, or cognizing), without any loss of force. We should still want to know why being unlearned and primitive implies the possession at birth of certain beliefs (or beliefs* or cognizings). Equivocation continues to dash Cummins's argument: only if he has in mind a standard meaning of 'belief' does the altered (P2) seem as obviously correct as he takes it to be, yet being primitive and unlearned only seems to stand any chance of implying the presence, at birth, of beliefs that mediate

architecture, then serves to explain how the requisite knowledge could be present at birth Thus, the idea that what is innate is known at birth motivates Cummins's brief discussion of what is implicit in the functional architecture of the cognitive system The same commitment explains the intuitive appeal of (P1) It seems plausible that a primitive LOT term's content must be fixed either by what is innate or what is learned (I say "it seems plausible," but I do not mean to endorse this intuition), if we then accept that a concept is innate because of what is known at birth about its application, it might seem that (P1) covers all the relevant possibilities

To dispel what intuitive appeal may still attach to (P1), I shall argue that it is unfounded to equate a concept's being innate with the newborn's possession of theoretical knowledge To this end, I now briefly survey some conceptions of innateness currently of interest to cognitive scientists, with the following question in mind To what extent does any one of these conceptions of innateness imply that, for a subject to possess an innate concept, she must possess knowledge at birth about that to which the concept applies? I need not survey every concept of innateness ever seriously proposed I merely intend to show that a good number of the conceptions of innateness presently on offer in cognitive science do not, taken together with CT and NTDP, sanction the inference from '*C* is innate' to 'Newborns possess knowledge of *C*'s application sufficient to facilitate reliable detection of that to which *C* applies' Thus, we can accommodate the common intuition that any given concept must be either innate or learned, without accepting that a concept's innateness implies the newborn's possession of the relevant kind of knowledge

Some of the conceptions of innateness I discuss imply that *something* relevant to concept acquisition is present in the newborn, and in some cases, what is present in the newborn seems similar to belief or knowledge (see footnote 19) But the degree of similarity falls short of what is needed to support Cummins's conclusion Although nativist views often impute knowledge or belief to the newborn, in many cases the mental states thereby attributed differ from knowledge and belief as these are normally conceived of, to such an extent that attributions of such states do not justify Cummins's combination of (P1) and (P2) Some special terminology will facilitate the discussion that follows Let me call *(P2)-knowledge* the kind of knowledge Cummins has in mind when he dismisses out of hand the possibility that

---

reliable detection, in keeping with (P1), if 'belief' is used in a nonstandard, permissive way newborns know enough about cats to facilitate the fixation of the content of 'cat' per CT. As it turns out, a number of prominent conceptions of innateness want identification of what is innate with anything that implies the presence of (P2)-knowledge at birth.

Consider first a pair of characterizations of innateness offered by Jeffrey L. Elman, Elizabeth Bates, Mark H. Johnson, Annette Karmiloff-Smith, Domenico Parisi, and Kim Plunkett[20]: "When we say that a form or a behavior is innate we really mean that, given normal developmental experiences, it is a highly probable outcome" (*ibid.*, p. 319; see also p. xiii). Note that (1) this conception of innateness is fairly attractive, for it seems to capture the main idea behind nativist arguments from the universality, or near universality, of outcome, while (2) this definition does not require that (P2)-knowledge be present in the newborn in order for some later-emerging behavior or cognitive structure to qualify as innate.[21] On this view of innateness, nothing more than the probability that a given outcome will appear in a given environment or range of environments determines the outcome's native, or nonnative, status. As a second option—the one they officially endorse—Elman et alia propose to use 'innate' "to refer to putative aspects of brain structure, cognition or behavior that are the product of interactions internal to the organism" (*op. cit.*, p. 23).[22] Here again, innateness does not require (P2)-knowledge. An innate concept must emerge from organismically internal relations, but this implies no particular form of knowledge present in the newborn supporting this emergence.

Elsewhere in developmental psychology, as well as in linguistics, much has been made of what appears to be an innate basis for a variety of capacities. It has been asserted, for example, that humans possess innate knowledge of grammar, an innate theory of mind, and an innate object concept, as well as innate domain-specific capacities for recognizing faces and distinguishing between phonemes[23] We do not find in this collection of work a distinctive or shared definition of "innateness", yet given the strong nativist claims often issuing from these quarters, it would be of interest if theorists working in these areas were to make nativist claims, while avoiding, or even eschewing, the attribution of (P2)-knowledge to the newborn I do not assert that no researcher working on these topics has ever claimed the existence of (P2)-knowledge at birth Nevertheless, even some of the staunchest nativists hedge their claims about innateness, as is illustrated by the two following examples

---

[20] *Rethinking Innateness: A Connectionist Perspective on Development* (Cambridge: MIT, 1996).

[21] It is a problem that discussions of innate knowledge often proceed absent any clear measure of what it is for knowledge to be explicitly represented, at birth or otherwise, and any precise view regarding to what extent and in what ways implicit representation can ground legitimate attributions of knowledge (or belief). Lacking such criteria, it becomes difficult to evaluate a claim that some bit of knowledge is or is not present at birth; see Clark, "What's Knowledge Anyway?" *Mind and Language*, xiii (1998): 571-75. Clearly, however, even if we count as knowledge the presence of whatever constraints lead to highly probably outcomes, this need not be (P2)-knowledge.

[22] The definition I initially quoted characterizes what Elman et alia sometimes call "primal" behaviors or cognitive forms.

First, take the case of Noam Chomsky,[24] probably the single most important intellectual force in the nativist revival He intentionally distances his nativist theses from what I am calling (P2)-knowledge While he claims that our knowledge of universal grammar is innate—it is "somehow represented in the genotype"—he is careful to remind us that 'knowledge', as he uses the term, does not necessarily imply knowledge as it is normally conceived of, to avoid confusion, he says that we "cognize" universal grammar (*ibid*, p 91)

Second, consider the case of the child's theory of mind While the theory of mind is frequently claimed to be innate, it is widely recognized that it takes three to four years for the theory to get up and running Very little, if any, of the theory is present at birth (the tendency to imitate, perhaps?) It seems to make perfect sense to many researchers working in this area to take a nativist position regarding the child's theory of mind without claiming that the newborn has (P2)-knowledge of the theory[25]

---

[23] There is a large literature addressing these issues For starters, see *Cognitive Science,* XIV (1990), which includes a series of papers by prominent developmental researchers investigating domain-specific theories and constraints, followed by Frank Keil's insightful metacommentary, also see various essays in Susan Carey and Rochel Gelman, eds , *The Epigenesis of Mind* (Hillsdale, NJ Erlbaum, 1991), Elman et alia (pp 107-08) list many of the domains in which innate knowledge or constraints have been thought to be operative and give references to much of the research taken to support nativist claims in these domains

[24] See, for example, *Rules and Representations* (New York Columbia, 1980), especially pp 69-70, 90-92 Here Chomsky expresses an ambivalent attitude toward knowledge as it is normally conceived of and its relation to cognizing Compare "In fact, I don't think that 'cognize' is very far from 'know' where the latter term is moderately clear " (p 70), and "In fact, it is not at all clear that the ordinary concept of 'knowledge' is even coherent " (p 92) (I do not mean to suggest that Chomsky's views are inconsistent, perhaps the ordinary concept of knowledge is incoherent overall, although there are certain limited uses of 'knowledge' that are "moderately clear ") Chomsky makes little effort to resolve his ambivalence because, as he sees matters, it is of no import to his nativist position whether cognizing and knowing (as it is normally conceived of) are the same mental-state type, or whether any sense can ultimately be made of the common conception of knowledge (p 92) For present purposes, the important point is this Chomsky does not think that a plausible nativist position in linguistics must imply the presence at birth of (P2)-knowledge of universal grammar

Two views of innateness remain, both of some stature The first is intended to capture the idea that genes determine innate traits, behaviors, or concepts Richard Samuels[26] expresses the view as follows "A characteristic C is genetically determined for an organism O just in case organisms that have the same combination of alleles as O develop C in all standard environments" (*ibid*, p 564) Without much comment, it should seem clear that this characterization of innateness is perfectly consistent with, and in fact seems specifically constructed to allow, the epigenetic emergence of innate characteristics it requires nothing in the way of knowledge, and certainly not (P2)-knowledge, in the newborn in order that a trait, behavior, or concept be innate

Of the forms of nativism discussed here, the final view to be considered might well be closest to what Cummins had in mind when formulating his critique of CT According to this view, a concept is innate if and only if it is primitive, where being primitive now means *not being acquired by hypothesis testing and confirmation*[27] Cummins's discussion of innateness contrasts those LOT terms relative to which explicitly represented theory mediates content fixation and LOT terms relative to which it does not By identifying innate terms with those in the latter class, Cummins seems to be following Jerry Fodor, for one might think there is little, if any, distance between LOT terms not acquired by the framing of explicit theory and those not acquired by hypothesis confirmation If I have accurately captured Cummins's thinking, however, he errs in his treatment of Fodorean primitives he gives no argument to show that a LOT term's being unlearned, in Fodor's sense of not being acquired by hypothesis confirmation, requires that the newborn have (P2)-knowledge about that to which the term applies

---

[25] Consider Stephen P Stich and Shaun Nichols's revealing remark about our innate knowledge of specific content domains, including our theory of mind "[I]t is plausible to suppose that natural selection has provided the child with lots of help—either in the form of innate knowledge structures or in the form of special purpose learning mechanisms"—Stich, *Deconstructing the Mind* (New York Oxford, 1996), p 151 This disjunctive offering suggests the extent to which the claim of an innate theory of mind is neutral in respect to the amount or kind of knowledge present at birth For an overview of research on the child's theory of mind, see Janet Wilde Astington, *The Child's Discovery of the Mind* (Cambridge Harvard, 1993), she notes the gap between nativism about the theory of mind and the stronger claim that the newborn *knows* the theory of mind (pp 167-68)

[26] "What Brains Won't Tell Us about the Mind A Critique of the Neurobiological Argument against Representational Nativism," *Mind and Language*, xiii (1998) 548-70

[27] Fodor has been known to frame the debate between nativists and empiricists as a disagreement about the number of primitive concepts, taking primitive concepts to be those not learned by hypothesis confirmation (in effect, those not reducible by definition to others), nativists claim there are lots, while empiricists claim that such concepts constitute a much smaller portion of our conceptual repertoire (the sensory concepts only)—"The Present Status of the Innateness Controversy," pp 281-83, 315

In fairness to Cummins, we should note the misleading nature of some of Fodor's own remarks on the acquisition of primitives In *The Language of Thought,* Fodor says "My view is that you can't learn a language unless you already *know* one" (*op cit*, p 65), going on to characterize the known language as innate This suggests the unnecessarily strong notion of innateness assumed by Cummins, according to which a concept's being innate implies the presence of certain bits of (P2)-knowledge at birth Nevertheless, as Fodor proceeds, he explicitly repudiates this strong notion of innateness "The view presently being proposed doesn't require that the innate conceptual system must literally be present 'at birth', only that it not be learned "[28]

The preceding survey, while necessarily brief, illustrates the genuine difficulty Cummins faces His argument against CT succeeds partly by taking an expansive view of the class of learned terms and painting as hopelessly inadequate the set of primitive, unlearned terms The latter set, Cummins implies, consists only of innate terms, about whose applications we must, at birth, possess knowledge sufficient to mediate content fixation, and clearly, we have very few of these But given that contemporary cognitive science embraces a wide range of conceptions of nativism, many of which do not require

---

[28] *The Language of Thought,* p 96 Fodor sometimes separates these issues to an even greater extent, so that a concept's being acquired by a process other than hypothesis testing—by brute-causal triggering, for example—does not itself imply the concept's innateness—see *Concepts Where Cognitive Science Went Wrong* (New York Oxford, 1998), p 137, also see Cowie, pp 104ff This is not merely a point about Fodor's intellectual development There is good reason not to infer a concept's innateness solely from the fact that the concept is acquired via a brute-causal process Consider Samuels's example of learning Latin by taking a Latin pill, which creates mastery of Latin in the consumer via a brute-causal process, just because Latin is thereby acquired hardly implies that knowledge of Latin was innate (or that the subject knew anything about Latin at birth) (*op cit*, p 554, note 6) (This would appear to be a variation on Danto's Spanish-pill example, which he uses to support the point that being unlearned does not imply being innate (*op cit*, p 128)) Note, though, that Fodor does not necessarily feel the weight of the preceding point, his reason not to infer innateness from brute-causal acquisition seems less robust, following more from considerations of what would be "a good thing tactically"—see his "Doing without What's Within Fiona Cowie's Critique of Nativism," *Mind,* xcix (2001) 99-148, p 131

(P2)-knowledge at birth, Cummins's approach seems misleading Even if we identify the unlearned with the innate—in keeping with the intuition that any given concept must be either innate or learned—innateness does not imply the presence of the relevant (P2)-knowledge

IV THE EMERGENCE OF LOT TERMS CHARACTERIZED NONSEMANTICALLY

Now consider the genesis of the vehicles of mental content As Cummins describes LOT, it must satisfy the following condition "It has a finite number of semantically primitive expressions *individuated syntactically* "[29] I shall construe the requirement of syntactic individuation as a requirement that LOT terms be individuated nonsemantically (there is, however, much more to be said about this issue[30])

---

[29] "The LOT," p 535, emphasis added

[30] Some of which I shall say here In characterizing LOT terms, we do best to observe the two following requirements, which neither singly nor jointly imply that *syntactic properties alone* serve to individuate LOT terms nonsemantically (1) in order that LOT be sufficiently language-like to be called a "language" (perhaps, for example, so that it manifests combinatorial rules supporting productivity), atomic units of LOT must be assigned to categories roughly analogous to the syntactic categories of natural language, and (2) for LOT to remain true to the quasi-reductive, naturalistic project it is normally taken to be part of, there must be *some* way of individuating LOT terms that does not depend on the terms' semantic contents Both constraints may be satisfied without its being the case that LOT terms are individuated solely by syntax (understanding syntax narrowly, to comprise only functional relationships holding among LOT terms) Say that LOT terms are individuated by reference to something other than, or more than, their syntactic roles, for example, by reference to neural types Condition (2) would clearly be met Also, in keeping with condition (1), each LOT term may well be of (at least) one determinate syntactic type (for example, noun, verb—or their LOT analogues), yet a term's having this, as well as other syntactic properties, need not constitute the term's identity conditions For discussion of syntactic individuation and related issues, see Michael Devitt, "A Narrow Representational Theory of the Mind," in William G Lycan, ed , *Mind and Cognition A Reader* (Cambridge Blackwell, 1990), pp 371-98

To pursue the argument of the present section, I need only assume the satisfaction of the weak condition that there should exist nonsemantic, although not necessarily syntactic, individuation criteria for LOT terms When the LOT theorist says, '*That item* has content *C*', I expect the LOT theorist to be able to describe what *that item* is, in some way other than by reference to *C* Note that this condition does not demand anything nearly as strong as might be demanded by a species-wide type-type identity criterion, which is often construed to require that a given LOT term individuated nonsemantically should be found in all the various members of the same species and found to have the same content in all members of the species In order that quasireductive cognitive theories invoking LOT possess the explanatory value they are supposed to, these theories would seem to have to assume no more than a certain amount of intrasubjective stability in the relationship between the subject's LOT terms individuated nonsemantically and the contents of those terms A satisfactory explanation of the way in which an individual subject's cognitive system implements a given content-based psychological generalization might

The question then naturally arises "What are the nonsemantic individuation criteria?" Although alternative approaches exist, those inclined toward realism regarding mental representations often assume that we can, in principle, state these criteria in the language of neuroscience [31] When we think of LOT terms in this way, we see that LOT terms, even those typically classified as innate, emerge in the child's cognitive system over time the very neural structures that are

---

appeal only to the repeated appearance of a vehicle possessing the same content *in that subject* on different occasions when the subject instantiates the regularity in question (the precise degree of intrasubjective stability of the vehicle required in order that these hoped-for explanatory purposes be served is an issue for another day) So long as contentful states can be attributed to a range of subjects, we can hope for content-based psychological generalizations that apply to all of those subjects, regardless of what variations may appear, from subject to subject, in the intrasubjectively stable vehicles to which those contents are attached

[31] This assumption is, of course, controversial For evidence of repeatable units that correspond to natural language terms and are susceptible to neurological characterization, see Friedemann Pulvermüller, "Words in the Brain's Language," *Behavioral and Brain Science*, XXII (1999) 253-79 For suggestions as to how mental representations corresponding to linguistic elements develop epigenetically and as to how the nature of developmental interactions affects the character of the nonsemantically individuated mental representations of words, see p 263, see also Ralph-Axel Müller, "Innateness, Autonomy, Universality? Neurobiological Approaches to Language," *Behavioral and Brain Sciences*, XIX (1996) 611-31, p 629 For recent attempts to defuse what have become common multiple realizability-based objections to neural individuation of LOT terms, see John Bickle, *Psychoneural Reduction* (Cambridge MIT, 1998), especially chapter 4, William Bechtel and Jennifer Mundale, "Multiple Realizability Revisited Linking Cognitive and Neural States," *Philosophy of Science*, LXVI (1999) 175-207, Robert W Batterman, "Multiple Realizability and Universality," *British Journal for the Philosophy of Science*, LI (2000) 115-45, also see Mark Wilson, "What Is This Thing Called "Pain"? The Philosophy of Science behind the Contemporary Debate," *Pacific Philosophical Quarterly*, LXVI (1985) 227-67 More general concerns about multiple realizability can be found in the essays in Part II of Jaegwon Kim, *Supervenience and Mind* (New York Cambridge, 1993)

One alternative is to identify the vehicles of content with vectors or regions in a state-space—see Evan Tiffany, "Semantics San Diego Style," this JOURNAL, XCVI, 8 (August 1999) 416-29, and my "On the Relationship between Naturalistic Semantics and Individuation Criteria for Terms in a Language of Thought," *Synthese*, CXVII (1998) 95-131 Were the LOT theorist to state term-individuation criteria in the language of state-spaces, and vectors and attractors therein, an epigenetic argument parallel to the one given in this section would seem to lead to the same conclusion reached in the text Even in cases of many concepts normally classified as innate, *experience shapes* the relevant attractors, neither the attractors nor their relevant ancestors are present at birth, and thus, neither would seem to be the contents of those attractors, or of the components thereof (but see footnotes 35 and 36 below)

It is sometimes suggested that we individuate LOT terms by nonsemantically specified, functional criteria In *Concepts*, Fodor would seem to have this approach in mind when he talks of modes of presentation (p 19) Individuation by functional role appears, however, to lead to serious problems for Fodor, as a LOT-cum-causal theorist—see Murat Aydede, "Fodor on Concepts and Frege Puzzles," *Pacific Philosophical Quarterly*, LXXIX (1998) 289-94 to become vehicles for content appear largely as a result of the child's interaction with her environment.

A summary excursion into developmental neurobiology may help to clarify my point. Developing brains grow and change in numerous and varied ways. Despite what looks to be a messy, nonlinear process of growth and change, stable neural structure emerges, which subserves the subject's representational capacities. In an attempt to understand this process of emergence, neuroscientists ask such questions as: (1) How does the overall functional structure of the brain appear without there being a tiny map of the brain's ultimate layout contained in the genetic material? (2) How does experience contribute to the development of the functional properties of various neural structures? A commonly given answer to the first question leads immediately to prosecution of the second. It is widely thought that overall structure emerges via the delicate interplay of multiple factors: there is, it is said, an epigenetic dance, experience appearing as one of the key participants. To a significant extent, the subject's experience shapes her neural substrate, determining the way in which the physical material of her brain eventually realizes its function as the center of cognitive processing. The interesting debate here is not over whether experience affects neural development—the experimental evidence seems to have settled that issue—but over the precise role of experience in shaping the structure of the brain.[32]

On one side in this debate stands the selectionist approach, a view closely associated with the work of Gerald Edelman.[33] According to selectionist views, genetically programmed growth provides a wealth of raw material that experience, among other factors, then winnows to create functional neural structure. Early development consists partly of the spreading of billions of dendritic and axonal arbors. Rich in synapses, these overlapping branches provide an enormous number, in the quadrillions, at least, of possible ways in which groups of neurons might excite or inhibit each other. Of these various patterns of connections, experience reinforces and strengthens those which have proven most useful to the organism, or which have simply been used more than competing ones. The activity of strengthened patterns of connectivity overshadows the less useful or less used patterns of connections, resulting in ineffective activity in the overwhelmed connections, dendritic or axonal retraction or withering, or even cell death

---

[32] See Patricia S. Churchland and Terrence J. Sejnowski, *The Computational Brain* (Cambridge: MIT, 1992), pp. 132, 307; D. Purves, G. J. Augustine, D. Fitzpatrick, L. C. Katz, A. S. LaMantia, and J. O. McNamara, eds., *Neuroscience* (Sunderland, MA: Sinauer, 1997), chapter 22.

[33] *Neural Darwinism: The Theory of Neuronal Group Selection* (New York: Basic Books, 1987); also see J. P. Changeaux, *Neuronal Man* (Princeton: University Press, 1983/85, second edition, 1997).

Constructivism offers an alternative[34] Constructivists agree with selectionists that experience contributes to regressive events They assign a further role to experience, however it guides the generation of axonal and dendritic arbors and can selectively increase the number of synapses in a given area of the brain or a given neural circuit, experience does so at least as early as the perinatal stage and as late as adulthood Constructivists claim that such environmental instruction causes the growth of the very neural resources that are needed to represent aspects of problems the subject will solve using those resources the specific character of experience catalyzes the growth of the neural resources needed to process more effectively the input that triggered the development of those resources While some selectionists accept that there are multiple waves of growth and pruning within the life span of one organism, this is a fundamental point of disagreement between the two camps whether experience can, as constructivists claim, cause axonal and dendritic growth or increases in the number of synapses in a way that is sensitive to the content or structure of the input itself

The preceding discussion of perspectives on neural development gives some indication of the prevalence of epigenetic views among contemporary neuroscientists[35] This point serves three purposes (1) to underscore the false nature of Cummins's dilemma offered in (P1), by describing some mechanisms likely to govern the emergence of nonsemantically individuated LOT terms in humans, mechanisms that would seem operative in cases of at least some primitive, unlearned LOT terms, (2) to illuminate the integrative vision above attributed to Pylyshyn, involving the purely physical emergence of new elements of cognitive architecture, which elements then subserve cognitive processes conceived of semantically, and (3) to provide background for the discussion of the following section, wherein I put to further use the distinction between LOT terms typed by their content and LOT terms conceived of nonsemantically.

---

[34] See, for example, Steven R Quartz and Sejnowski, "The Neural Basis of Cognitive Development A Constructivist Manifesto," *Behavioral and Brain Sciences*, xx (1997) 537-96 (page numbers include peer commentary)

[35] One might take exception to this characterization of the state of developmental neurobiology, on the ground that selectionism seems to support a fairly strong nativism (see, for example, M Piatelli-Palmarini, "Evolution, Selection, and Cognition From 'Learning' to Parameter Setting in Biology and in the Study of Language," *Cognition*, xxxi (1989) 1-44) Does not selectionism imply the presence at birth of a surfeit of nonsemantically individuated LOT terms, the bulk of which later disappear in a manner that accords with selectionist principles, the remaining terms having been "present from birth?" This would, however, misrepresent what is plausible in selectionist thinking, for first, a significant percentage of the human brain's synaptic connections develop after birth, this is a widely accepted view with which selectionists should take no particular exception—see Elman et alia, pp 291-94, for selectionist endorsement of this view, see Changeaux, p 248 (second edition), second, in most cases, experience-based reinforcement and pruning are required before whatever patterns of neural connections that exist at birth cohere as functional units Thus, if we identify LOT terms, thought of nonsemantically, with functional cell assemblies, selectionism does not support the radically nativist view of nonsemantically individuated LOT terms It does not imply that all, or even most, of the adult's nonsemantically individuated, primitive, unlearned LOT terms were present at birth

Regarding any given primitive LOT term, Cummins asserts that the causal theorist must choose between its being (a) innate, the fixation of the term's content mediated by knowledge possessed at birth, and (b) acquired at least partly by the subject's formulating an explicit theory of that to which the term applies. If selectionists *or* constructivists are correct, the vast majority, perhaps all, of an adult's nonsemantically individuated LOT terms appear in no substantial form at birth. This list of nonsemantically individuated concepts not present at birth seems quite likely to include such LOT terms as 'physical object', 'subject' (in the grammatical sense), and 'belief', terms that would seem primitive and unlearned and, when categorized according to their content, are often considered innate. Many such concepts *are* plausibly innate, in some sense: they emerge in all standard subjects across a wide variety of contexts. Nonetheless, they do not fit Cummins's description of unlearned terms. They do not seem to be associated with content-fixing knowledge present at birth, even implicitly represented knowledge. How could they be? These concepts, nonsemantically individuated, have yet to appear on the developing subject's cognitive scene.[36] Thus, such terms do not seem to fall into either of the categories Cummins recognizes The causal theorist should feel free to account for the fixation of these terms' contents by appeal to something other than the processes allowed by (P1) [37]

---

[36] It is possible that content transfers from one nonsemantically characterized vehicle to another throughout the course of development (also, because plausible identity conditions for individual vehicles should allow changes over time in important physical properties of what continues to count as the same nonsemantically individuated unit, there may be some as yet unspecified sense in which the vehicle is present at birth—but see the preceding note). Perhaps the content of an innate concept, present at birth, attaches to one nonsemantically individuated LOT term, then "hops" to another as the next "host" develops. Recognizing this possibility, though, hardly makes it probable that there are present at birth vehicles carrying *all* the contents of terms typically claimed by cognitive scientists to be innate (or, more generally, of *all* of the primitive, unlearned LOT terms).

We should also consider the possibility that one could possess at birth knowledge regarding the application of a LOT term even when that term has not yet formed in one's cognitive or biological systems, however, without further development of this proposal, it is difficult to say whether it is at all plausible in the human case.

V LOT TERM INTRODUCTION AND THE CONTENT OF LEARNED TERMS

For a primitive concept to be unlearned, the newborn need not possess a content-determining theory of that to which the term applies, furthermore, she need not possess the nonsemantically individuated LOT term that will one day carry the content by reference to which we would normally identify that concept Or so I have argued [38] Nonetheless, what has been said so far does not show that any particular LOT term, such as 'cat', is unlearned, nor does it show that all LOT primitives are unlearned Consequently, to round out my defense of CT, I must now address the difficult case, where explicitly represented theory seems necessary to mediate content-fixing reliable detection

---

[37] Here Cummins may wish to follow some prominent nativists, dismissing the present discussion of neurobiological issues because of the paucity, to date, of substantive results in cognitive neuroscience In Fodor's review of *Rethinking Innateness,* he lambastes Elman et alia for assuming that certain neural mechanisms must encode representation "[T]here isn't one, *not one,* instance where it's known what pattern of neural connectivity realizes a certain cognitive content, innate *or* learned, in either the infant's nervous system or the adult's " (*The London Review of Books,* xx, 12 (1997)—reprinted as Chapter 12 of Fodor's *In Critical Condition Polemical Essays on Cognitive Science and the Philosophy of Mind* (Cambridge MIT, 1998), pp 143-51, the quoted passage appears on p 145) Chomsky expresses a similar view when resisting the demand that we cash out in neural terms nativist claims in linguistics, see "Language and Nature," *Mind,* civ (1995) 1-61 According to Chomsky, "The current situation is that we have good and improving theories of some aspects of language and mind, but only rudimentary ideas about the relation of any of this to the brain" (p 11) He feels free, then, to make strong nativist claims in the absence of any supporting theory of neural implementation "The conditions of language acquisition make it plain that the process must be largely inner-directed, as in other aspects of growth, which means that all languages must be close to identical, largely fixed by initial state" (p 17) Compelling considerations, including poverty of stimulus arguments, imply innate content, it matters not whether we understand how precisely that arises from the subject's underlying physical being

It seems to me that Fodor and Chomsky simply refuse to acknowledge recent decades' progress in cognitive neuroscience, although this is not the place to support my claim To the extent that my claim is correct, however, Cummins should want to avoid endorsing Fodor's and Chomsky's positions *vis-à-vis* the relevance of neuroscience to our evaluation of concept nativism Furthermore, the arguments of the preceding and following sections stand independent of any controversial claims about the relation between cognitive activity and development and their neurological bases

[38] In presenting these arguments, it has been convenient to separate discussion of the emergence of content and questions about the emergence of the vehicles of content The causal theorist might well suspect, however, that these two processes are closely connected In many cases, the very causal interactions with the environment that shape the brain's nonsemantically characterized representational resources at the same time contribute significantly to the fixation of the content of the units so shaped, so that, to a great extent, what does the shaping often exemplifies the kind to whose members the shaped entity comes to refer—see my *op cit*, and Müller, p 629 Adopting this more integrated view of conceptual development would in no way weaken the arguments in the text, for the integrated view allows concepts to emerge in the dialectically pertinent ways they can be unlearned,

Assume, then, plausibly enough, that the explicit formulation of theory mediates reliable detection for 'cat', via such theoretical axioms as Cummins lists "(A) Cats have whiskers (B) Cats have four legs (C) Cats have fur "[39] Is there a way for the subject to frame such explicitly represented axioms without requiring—absurdly, if we accept CT—that the subject already possess a concept with the content *cat*?

This is possible, if we envision the initial theoretical axioms as containing the nonsemantically individuated term 'cat', as opposed to a LOT term with the content *cat* In section IV, I described some of the ways in which LOT terms, conceived of nonsemantically, might be introduced at the neural level early human development comprises an ongoing process of neural generation, growth, and death and synaptic generation, strengthening, and weakening which results in the emergence over time of functionally coherent cell assemblies If this picture is correct, then at least some of the time after birth, the human brain/cognitive system generates, or 'coins', new LOT terms, individuated nonsemantically, it generates new circuits or cell assemblies of the same sort as those we describe when we characterize LOT terms in the language of the neurosciences It is possible, then, that the subject's early theory of catness include a newly coined LOT term lacking the content *cat*, perhaps lacking content altogether If this picture is correct, then the initial theoretical axioms that include the nonsemantically individuated LOT term we call 'cat' would more properly be thought of in the following way (A) *t* (a newly coined, nonsemantically individuated LOT term) have whiskers, (B) *t* have four legs, (C) *t* have fur [40]

---

perhaps even innate, emerging via postnatal interaction with the environment, all without the newborn's possessing (P2)-knowledge

[39] "The LOT," p 537

[40] How might such terms be coined? Changeaux suggests that protorepresentations (or something much like them—"prerepresentations," he calls them) briefly cohere and pass away due to spontaneous, oscillatory firing of neurons (*op cit*, pp 139, 145, 169), experience with the world can then cause increased stabilization of certain prerepresentations, in keeping with the selectionist view summarized above

Research that takes a dynamical systems-based approach to cognitive development suggests an alternative working within the framework provided by dynamical systems theory, Esther Thelen and Linda Smith—*A Dynamic Systems Approach to the Development of Cognition and Action* (Cambridge MIT, 1994), p 134—propose that the child harnesses, by a process similar to trial and error, naturally generated

Cummins's fundamental objection to CT is thus dissolved. Take learned concept *C*. Given CT and NTDP, the subject cannot possess

---

"noise" to create new solutions, sometimes fairly stable ones, to the problems of cognitive and motor development. Although Thelen and Smith would almost certainly chafe at my referring to the harnessed units as 'terms in LOT', their view may in fact explain how the brain coins LOT terms, by transforming snippets of relatively chaotic processes into stable attractors in a dynamical system.

There are two general ways in which we can conceive of the semantic status of such LOT terms as 'fur', 'legs', and 'whiskers', as these terms appear in the axioms associated with 'cat' (call these *auxiliary terms*): either their content is in place at the time the axioms are framed, or they, like 'cat,' initially appear in the axioms as protorepresentations. When, in the text, I discuss explicitly represented theoretical axioms, I take all auxiliary terms to have their content fixed. One reader expressed concern that this assumption entails an unacceptable regress. Thus, in what follows, I explain in more detail the semantic status of auxiliary terms, and why their role in the coining of new LOT terms is unproblematic.

There are three routes by which auxiliary terms may have their content fixed prior to their appearing in the relevant axioms: either (1) by definitional combination of other LOT terms whose contents are fixed, (2) by the mediation of explicitly represented, but nondefinitional axioms in which auxiliary terms appear—axioms of just the sort presently under discussion, but directed at protorepresentations 'legs' or 'whiskers' instead of the protorepresentation 'cat', or (3) by the mediation of implicit theory. Following Cummins ("The LOT," p. 537), and in keeping with a broad consensus among philosophers of mind and language, I do not hope for much mileage from the first option. Definitions are too few, the case of 'bachelor' notwithstanding, and insofar as auxiliary terms sometimes are fully definable, we must explain how the contents of their definitional constituents were fixed. Eventually such definitional chains must be grounded in accordance with at least one of the other alternatives, (2) and (3). It is also possible that a given auxiliary term has its own set of explicitly represented, detection-mediating axioms, operative in the introduction of the auxiliary term itself. Of course, these axioms will contain their own auxiliary terms (for example, auxiliary terms relative to 'whiskers'—such LOT terms as 'long' and 'thin'). Thus, as in the case of the first option, this story cannot be complete: it does not account for the ultimate introduction of LOT terms. For this reason, case (3) is of fundamental importance. Cummins allows implicit theory to play a detection-mediating role for at least some LOT terms, and the auxiliary terms 'fur', 'whiskers', and 'legs' may well be such terms (and if they are not, then other relevant terms will be either the terms that make up their definitions, or the further auxiliary terms that appear in 'whiskers''s or 'legs''s explicitly represented, but nondefinitional, introducing axioms, and so on, as each case may be). It is not implausible to claim a significant stock of terms whose contents are fixed in accordance with possibility (3), to show this was part of the burden of the preceding sections. Surely, this stock is not rich enough to serve as a definitional base for all other concepts, but that is not what we are after. We wish to find enough such terms to provide definitional constituents in cases where newly introduced LOT terms have definitions, but also, and this is surely the more common case, to provide the materials on which to draw when framing nondefinitional axioms that mediate the tokening of newly coined LOT terms. Recursive applications of such a process can yield a wide range of learned, primitive, nondefined LOT terms from a relatively small base of terms whose content is fixed via the mediation of implicit theory.

Rather than handle the threat of regress in the manner just described, a reader suggests treating auxiliary terms as protorepresentations *at the time they appear in the* a concept with the content *C* unless she possesses a nonsemantically individuated vehicle '*C*', the vehicle must be present to hold a place in the subject's theory of *C*s But '*C*', characterized nonsemantically, does not have to have content *C* when it first appears in the axioms constituting the theory that mediates the fixing of content *C* to vehicle '*C*' The alleged circularity of CT vanishes

To flesh out and defend this proposal, I respond below to a series of critical questions

Question #1 What causes the brain/cognitive system to produce a new LOT term that will be the subject of an explicit theory?

Answer #1 When the subject encounters a cluster of properties not sufficiently similar to any stored cluster (each cluster embodying a mediating theory), her cognitive system introduces a new primitive term The term-introduction rule could itself be either explicitly represented or implicit in the architecture (although in what follows, I assume what I take to be the more plausible possibility that the rule is implicit in the system's workings) Of course, the system must already be capable of representing the properties in the cluster, but this does *not* imply that said cluster defines the newly introduced term (see footnote 40)

Question #2 When a new LOT term is introduced, why is content not fixed immediately and permanently?

Answer #2 Whether content is fixed immediately depends on (1) the mediating theory of that to which the newly introduced LOT term applies and (2) the exact nature of the causal relation supposed to underlie reference Even if content is fixed straightway, the LOT term in question, individuated nonsemantically, is not permanently

---

axioms in which such terms play their auxiliary role I do not rule out such a possibility It may be that as a result of the kinds of process stressed above—maturation, dendritic arborization in response to environmental stimuli, and so on—an axiom consisting wholly in protorepresentations may take shape in the subject But if other protorepresentations wholly mediate the context-fixing detection relations of a given protorepresentation *t*, the relevant axioms have the status only of an implicit theory of the things to which *t* will come to refer, for this is the dimension according to which we presently categorize a theory of *t* as explicit or implicit whether the reliable tokening of *t* in response to the relevant distal properties or objects is mediated by explicit representations In the present section, I set aside cases where mediating theory is implicit, including cases where axioms consist wholly of protorepresentations, for here I intend to show how a LOT term can be acquired by the mediating effects of *explicit* theory

It is possible that at least one of protorepresentation *t*'s axioms is partly explicit, partly implicit, containing at least one auxiliary term with content fixed and at least one protorepresentation beyond *t* I trust that the views presented here provide a framework for handling such mixed cases, but a full discussion of them would lead us further afield than constraints of space allow saddled with that content Imagine that the subject encounters a cluster of properties sufficiently unlike those encountered before, and in response, the cognitive system frames theoretical axioms that include a newly coined LOT term *t* Assume these axioms to play a sufficient mediating role that, according to a version of CT we have adopted, *t* acquires content The subject may later interact with a new item differing substantially from the item that caused the original coining of *t*, *without the new item's differing enough from the original to cause the coining of a new LOT term* The cognitive system responds by altering existing theoretical axioms or adding new axioms including *t* Whether this refinement changes the content of *t* depends on the causal relation established by this refinement, and this depends on the details of the case

Note that, if a new experience causes a change in theoretical axioms in which *t* appears, this does not imply that the cognitive system explicitly represented *error* in its earlier theory of *t* Whatever theory, implicit or otherwise, governs the introduction of new LOT terms may do so by measuring the match between a newly encountered item and stored, associated clusters of properties (such clusters are sometimes called "prototypes") In a case where the mismatch between the new item and stored clusters is not so great as to cause the coining of a new LOT term, the result is a value-neutral integration of the profile of the newly encountered item and whatever existing prototype most closely matches the new item's profile Pace Cummins,[41] then, it is conceivable that the cognitive system alter a prototype, which embodies a theory, without the cognitive system's having to represent the prototype, prior to its alteration, as normatively mistaken Implicit in the workings of the system is a theory that causes the coining of a new LOT term under some circumstances and the integration of new experiences and existing prototypes under others [42] Admittedly, such an implicit theory has peculiar status, perhaps that of metatheory Better still, we may think of the theory as part of the system's program, emerging from the arrangement of the physical matter that constitutes the cognitive system, in this case, the implicit "theory" may not be found among the cognitive system's mental states as we would normally describe them, for this reason, it may be misleading to call it a 'theory'.[43]

---

[41] "The LOT," p 538

[42] Here one might imagine a best-match, or 'nearest neighbor', system with a threshold of difference beyond which no match will be made, and instead a new term coined, furthermore, the system could be such that when a match is made and the stimulus differs from the prototype, the stimulus "pulls" the prototype toward it in similarity space For discussion of nearest neighbor, prototype-based categorization, see Churchland and Sejnowski, pp 76ff

Question #3. The model of concept acquisition I have suggested seems to allow for computation without representation, newly coined LOT terms would seem to be in the domain of some computational operations, despite the possibility that they represent nothing. Does not the orthodox computational approach to explaining cognition proscribe such models?[44]

Answer #3. Three points in response. First, it seems empirically unproductive to insist that the following holds for *any* physical state with an assigned value that falls in the domain of *any* formal operation in a computational model of *any* aspect of cognition: that physical state must itself have representational content. Computational models of stereopsis, for example, explain how the brain might construct a coherent representation of the immediate environment from two separate two-dimensional arrays, one originating at each retina. Such models do so by, among other things, assigning values to physical states and placing these values in the domain of the representation-constructing algorithms specified by the relevant computational theory of vision[45]

---

[43] That is, it may be a mistake to think that the cognitive system holds a theory as to when it will introduce a new LOT term (individuated nonsemantically) and when instead it will integrate the information given in new experiences into existing representational structures. In contrast, *I* surely am suggesting a theory (or, more accurately, a theory-outline) of the workings of one aspect of the cognitive system, see Cummins's discussion of this distinction in "Inexplicit Information," pp. 123-25.

[44] For statements of the orthodox view, see Fodor, *The Language of Thought*, pp. 31, 46, 56, 99, also see his *The Modularity of Mind* (Cambridge: MIT, 1981), p. 5, and *Concepts*, pp. 10-11, Pylyshyn, *Computation*, pp. 51-58, Churchland and Sejnowski, pp. 62-66. Whether the orthodox view grounds a criticism of the present proposal depends on how tightly one associates computation and semantically interpretable states. That is, one might claim that it is *essential* to computation that it only be carried out over semantically interpreted states or units, this understanding of computation seems to conflict with the present proposal. On the other hand, one might, as Pylyshyn does, think that our best cognitive theories will sometimes deal in computations over interpreted states, while at the same time including nonsemantic explanations of important aspects of cognition: "[W]e may need the biological level to explain other things as well, such as possibly the nature of cognitive development or maturation or psychopathology, and perhaps *some changes that are now called learning*, exactly what facts fall at each of the three levels [semantic, syntactic, and biological] remains to a large extent an open empirical question"—"Computing in Cognitive Science," in Michael I. Posner, ed., *Foundations of Cognitive Science* (Cambridge: MIT, 1989), pp. 51-91, p. 61, italics added, also see p. 72. I would extend Pylyshyn's remarks and claim that what sorts of state can be included in a computational model is also an empirical issue, if our most empirically successful computational models of cognition include dummy terms, with syntactic and biological properties, but no semantical ones, then so be it.

There is no reason, however, to think that every such physical state is itself a representation, granted, each such state is assigned a numerical value, but this does not imply that the physical state *represents* that numerical value Whether we assign representational status to such physical states depends on the explanatory value of doing so, and sometimes there seems to be little value in such making such assignments[46]

Second, I do not hypothesize the existence of a computational system containing no representations In the kind of system I have been discussing, representations abound, nevertheless, some newly coined LOT terms might not be among them Note that many of the advocates of content-laden interpretations of computational processes rest their advocacy largely on the empirical fruitfulness of semantic talk Such theorists typically agree that computational processes *can be* characterized in nonsemantic terms It is not incoherent to talk about computational processes in, say, syntactic terms, such theorists will admit, they argue that semantic interpretation is indispensable if our cognitive models are fully to explain human behavior or cognitive processing[47] The present view concurs to a great extent I do not contend the claim that many, perhaps the majority of, elements participating in the computational processes underlying human cognition possess semantic values Turnabout is fair play, however, and if the serving of empirical virtues sometimes favors nonsemantic descriptions (or as I have suggested, "mixed" descriptions), then we should accept such descriptions In particular, if we best explain the emergence of representational primitives by hypothesizing the presence of axioms, subject to computational manipulation, containing both representing and nonrepresenting elements, then this approach should be acceptable even to those, including myself, who think that content-based characterizations of some processes are required if we are fully to understand those processes

---

[45] On computational models of stereopsis, see David Marr, *Vision* (New York Freeman, 1982), pp 111-59, and Churchland and Sejnowski, pp 199-216

[46] For a similar point, see Frances Egan, "In Defence of Narrow Mindedness," *Mind and Language*, xiv (1999) 177-94, pp 184-86, but note that she takes a more pessimistic overall view of the importance of content ascription than is assumed here See also Chomsky's discussion of internalist computational views in "Language and Nature," especially pp 51-56

[47] Christopher Peacocke, "Computation as Involving Content A Response to Egan," *Mind and Language*, xiv (1999) 195-202, pp 198-99, Fodor, *Representations*, pp 22-23 Regarding the present proposal, it is unclear why any explanatory power is lost when we allow that some newly coined LOT terms participate in computations even though the terms lack content This suggests a condition that must be met by a causal theory of content it must assign content to newly coined, content-lacking LOT terms in a timely enough fashion that such terms possess content when our explanatory purposes demand that they do

Here it is also important to bear in mind that newly coined LOT terms are, presumably, of the sort of structure whose instances often have content in the computationally conceived-of human cognitive system On this view, all states that participate in cognitively relevant computational processes share certain non-content-determining, nonsemantically characterized traits, newly coined LOT terms, perhaps devoid of content, exhibit these traits and, as such, are candidates for participation in computational processes Recalling that computation can be described at a nonsemantic level, we should not automatically exclude these structures from participation in computational processing, instead we should see them as potential representations, which have not yet, but may eventually, come into content-conferring relations to kinds or properties It is in this sense that newly coined LOT terms can be protorepresentations, appearing in axioms upon which computational processes can operate, even when the content of the incipient representations is not yet fixed These protorepresentations may bear the physical marks of computed-upon representations before such protorepresentations fully come into the causal relation that will, according to CT, fix their content [48]

Lastly, allowing *some* computation without representation does not imply an unacceptable proliferation of cognition, the conditions for being a cognitive system are much more stringent than simply that a system sometimes compute something 'No computation without representation' is a fine slogan when understood properly nothing described as a computational system counts as a cognitive system unless it exhibits the kind of behavior or capacities best explained by positing internal representations over which at least some of the computations operate. So far as can be presently told, few, if any, nonhuman systems meet this condition.[49]

---

[48] It is sometimes claimed that semantics is prior to syntax for the following reason what counts as a syntactic property in LOT is determined by the physical (or otherwise nonsemantic) properties causally operative in cases where we interpret the cognitive system in semantic terms—see, for example, Tim Crane, "The Language of Thought No Syntax Without Semantics," *Mind and Language*, v (1990) 187-212 On this view, we must first characterize thought and behavior in semantic terms, then we can look for underlying formal or syntactic properties of, for example, bits of the brain, which will explain the occurrences of semantically interpreted behavior Here I take no issue with this view, for it is consistent with the present proposal Once we have discovered the operative formal properties—those which explain cognitively characterized skills and capacities—we are free to identify further instantiations of these formal properties as LOT terms, whether or not these terms yet play semantically characterized roles in the cognitive system

VI. CONCLUSION

If CT is to maintain any plausibility as a theory of content for human concepts, it must respect the known facts of human psychology, including human perceptual psychology. Cummins errs, however, in thinking that the demand for psychological realism supports his case against CT. To show this, I first considered the ways in which concepts might emerge in the cognitive system. Cummins sees only two options: for any given concept, the causal-cum-LOT theorist must choose between saying (a) the subject has knowledge, at birth, of that to which the concept applies and (b) the concept's content is fixed via the mediating effects of an acquired, explicitly represented theory of that to which the concept applies. I have argued that this range of options is artificially limited: we should recognize the wide range of cases in which mental content and the vehicles of mental content emerge after birth in the absence of explicit theory. Cummins admits the power of implicit theory to mediate content-fixing reliable detection in the case of concepts about whose applications we possess knowledge at birth, yet it appears that implicit theory's detection-mediating power extends well beyond these cases. Correlatively, we should doubt that explicitly represented theory participates in content fixation in as many cases as it has often been thought to. This point weakens Cummins's argument against CT, for the argument's force rests, to a significant extent, on the assumption that explicit-theory formation is a necessary condition of content fixation for a large class of human concepts.

More damning to Cummins's case against CT was the argument of the penultimate section. Even where explicit theory contributes to the fixation of content, there exists a possible route to content fixation within the framework assumed by CT, a route that Cummins fails to consider. It is agreed on all sides that the LOT theorist must individuate LOT terms nonsemantically, plausible ways of doing so allow us to see how a LOT term can appear, without content, in the framing of a theoretical axiom. This possibility defeats Cummins's claim that where explicitly represented theory mediates reliable detection, CT is trapped in an inescapable circle. For the time being, then, the causal theory of mental content stands as a live option.[50]

---

[49] One might wonder whether the present defense of CT is consistent with the nativism of one of CT's most ardent advocates: Fodor. Throughout, I have emphasized ways in which the subject's interaction with her environment shapes her cognitive resources, and thus the views presented here may seem squarely nonnativist. The apparently nonnativist implications of the present proposal may be appearance only, however, insofar as Fodor counts a concept as innate if and only if that concept is not acquired by a rational-causal process, the present view might in fact be consistent with Fodor's extreme nativism. Much depends on how one conceives of the axioms discussed above, whether, for example, they consist in hypotheses about the meaning of a newly coined LOT term, or whether their constituent auxiliary terms play only a brute-causal, triggering role. No doubt we should like to know (a) whether the present view implies the innateness of any LOT terms—by a Fodorean criterion or any other prominent one—and (b) if so, which LOT terms would thereby be classified as innate. Nevertheless, given the complexity of such issues, limitations of space prevent their satisfactory prosecution here.

ROBERT D. RUPERT

Texas Tech University

---

[50] More than one reader has raised the concern that my defense of CT coheres poorly with an atomistic version of CT, such as Fodor's. Because a newly coined protorepresentation appears in axioms alongside full-fledged representations, one might think that, even though the axioms do not provide definitions, the new LOT term inherits its content from its relations to these full-fledged representations. Bear in mind two points. First, Fodor allows intentional mechanisms to mediate the causal relations that atomistically fix content for LOT terms, atomism is preserved because the ultimate ground of content remains the causal relations themselves, nonsemantically characterized (*Psychosemantics*, p. 121, and *Concepts*, pp. 78-79), in fact, Fodor allows that in the human case, there may be certain concepts we cannot acquire without mediation by intentional mechanisms, but he allows this while still adhering, at a more theoretical level, to his semantic atomism (*Concepts*, pp. 157-58). For further discussion of these points, see my "Dispositions Indisposed: Semantic Atomism and Fodor's Theory of Content," *Pacific Philosophical Quarterly*, LXXXI (2000): 325-48. Second, in defending CT here, I intend to be defending, if only indirectly, the version of CT I have proposed elsewhere—"On the Relationship between Naturalistic Semantics and Individuation Criteria for Terms in a Language of Thought", "The Best Test Theory of Extension: First Principle(s)," *Mind and Language*, XIV (1999): 321-55, and "The Best Test Theory of Extension" (Ph.D. Dissertation, University of Illinois at Chicago, 1996). I advocate a hierarchical version of CT, according to which, at the lowest level, content is fixed atomistically, with a certain degree of molecular content-fixation allowed at higher levels. Regarding the present concern, that mediation by intentional mechanisms fixes content nonatomistically, I would, at least with respect to many learned, primitive LOT terms, follow Fodor, emphasizing the first of the two points made above. In certain cases, however, I incline toward allowing a greater degree of holistic (or, more precisely, molecular) content-fixation than does Fodor, and thus, I see no need to show that my proposals here support a purely atomistic CT. Limitations of space require that I omit detailed discussion of the hierarchical view and my reasons for holding it, as opposed to a strictly atomistic version of CT.

ALTER EGOS AND THEIR NAMES*

Failure of substitutivity of coreferential terms, one of the hallmarks of referential opacity, is standardly explained in terms of the presence of an expression (such as a verb of propositional attitude, a modal adverb, or quotation marks) with opacity-inducing properties It is thus assumed that any term in a complex expression for which substitutivity fails will be within the scope of an expression of one of these types, and that where there is an expression of one of these types there will be failure of substitutivity for terms within its scope

I shall discuss a series of examples that have been thought to challenge this explanation by exhibiting failure of substitutivity of coreferential terms for positions not within the scope of any of the standard opacity-inducing expressions If these examples are genuine, then the usual explanations of opacity are either incomplete because there are sources of opacity other than those standardly identified, or completely mistaken because the standardly identified expressions are not causes of opacity I shall argue, however, that the examples exhibit only failure of substitutivity of *non*-coreferential terms, and, hence, do not present a challenge to standard explanations of opacity

I SAUL'S PUZZLE

Jennifer Saul[1] has drawn attention to sentences containing no propositional-attitude, modal, or quotational constructions that exhibit apparent failure of substitutivity of coreferential terms For example, whereas (1) and (2) are true, (1′) and (2′) are not[2]

---

* I have benefited in the writing of this article from conversations with Robert M Adams, Brad Armour-Garb, Andy Egan, Bob Fiengo, Fred Goldstein, Michael Huemer, Jerry Katz, Steve Leeds, Eric Saidel, Jennifer Saul, Virginia Valian, and Mark van Roojen Special thanks are due to Edward Becker and George Bealer, with whom I discussed the issues dealt with here at length, and who offered penetrating criticism and helpful suggestions on earlier drafts

[1] "Substitution and Simple Sentences," *Analysis*, LVII (1997) 102-08, "Reply to Forbes," *Analysis*, LVII (1997) 114-18, "Substitution, Simple Sentences, and Sex Scandals," *Analysis*, LIX (1999) 106-12

[2] These examples are mine, (3) and (3′) are Saul's She gives a number of examples, including

(i) Clark Kent went into the phone booth and Superman came out

(i′) Clark Kent went into the phone booth and Clark Kent came out

and

(ii) Superman leaps tall buildings more often than Clark Kent

(ii′) Superman leaps tall buildings more often than Superman

0022-362X/01/9810/531–52

(1) Batman is Bruce Wayne in disguise
(1′) Bruce Wayne is Bruce Wayne in disguise
(2) Clark Kent is Superman's secret identity
(2′) Superman is Superman's secret identity

Moreover, the truth of (3)

(3) Clark Kent always arrived at the scene just after one of Superman's daring rescues

does not guarantee the truth of (3′)

(3′) Superman always arrived at the scene just after one of Clark Kent's daring rescues

Let us call this situation—that there are apparent failures of substitutivity in linguistic contexts containing no patent source of opacity—*Saul's puzzle*

Saul argues that there is no obvious resolution of her puzzle consistent with standard accounts of the semantics of opaque contexts It cannot be resolved by a straightforward Fregean account, on which 'Superman' and 'Clark Kent' would be taken to refer to their senses in the contexts (1)-(3′), since senses do not wear disguises or have secret identities or arrive at scenes [3] Metalinguistic accounts also fail (4) and (4′), for example, are just as puzzling as (1) and (1′)

(4) The bearer of 'Batman' is the bearer of 'Bruce Wayne' in disguise
(4′) The bearer of 'Bruce Wayne' is the bearer of 'Bruce Wayne' in disguise

Saul also considers a solution on which (1′) and (2′) are true, (3) entails (3′), (ii) is false, and (i) entails (i′), and intuitions to the contrary are explained away on the basis of pragmatic considerations (For example, as due to our acceptance of the (in these cases false) implicature that the distinct names in (1), (2), (3), (i), and (ii) refer to distinct things (Why else would different names be used in these sentences?))[4] Although perhaps not everyone will accept Saul's intuitions about these cases, they strike me as essentially sound I am not, therefore, prepared to sacrifice the falsity of (1′) and (2′), the nonentailment of (3′) by (3) and (i′) by (i), and the truth of (ii) to a pragmatics-based solution.

---

[3] "Substitution and Simple Sentences," p 104

[4] Other solutions that Saul discusses and rejects will be discussed below For further debate, see Graeme Forbes, "How Much Substitutivity?" *Analysis*, LVII (1997) 109-13, and "Enlightened Semantics for Simple Sentences," *Analysis*, LIX (1999) 86-91, Joseph G Moore, "Saving Substitutivity in Simple Sentences," *Analysis*, LIX (1999) 91-105, and Stephen Predelli, "Saul, Salmon, and Superman," *Analysis*, LIX (1999) 113-16

Moreover, I do not share Saul's puzzlement about these cases, for it seems to me that the most straightforward explanation of the substitution failures—namely, that 'Superman' and 'Clark Kent', 'Bruce Wayne' and 'Batman' are *not* coreferential—is correct. Saul considers this sort of solution, but rejects it on the grounds that it entails that (5) and (6) are false.

(5) Superman is Clark Kent.
(6) Batman is Bruce Wayne.

It is, for Saul, a condition of adequacy on any solution to her puzzle that it preserve the truth of (5) and (6).[5]

There are, however, strong motivations for defending this solution to Saul's puzzle, in spite of the fact that it entails the falsity of (5) and (6). In addition to the effect on traditional explanations of opacity mentioned above, there is the further concern that, given that opaque contexts are those in which substitution of coreferential expressions can fail, Saul's examples seem to show that virtually *all* contexts are opaque.[6]

Of course, if the intuition that (5) and (6) are literally true really is non-negotiable, as Saul claims, then we really might have a puzzle. But I think we can sustain the thesis that (5) and (6) are literally false, and explain away intuitions to the contrary on the grounds that sentences like (5) and (6) have fairly common nonliteral uses on which they express closely related propositions that may be true.

I shall argue that 'Superman' and 'Clark Kent' are names of *alter egos*, that alter egos are distinct from the persons whose alter egos they are, and that utterances of identity sentences containing one or more occurrences of an alter-ego name should, if they are to be taken to be true, be interpreted as meaning either that the person named by one term of the identity is the person whose alter ego is named by the other, or (as in the case of (5), see below) that the alter egos named by the terms of the identity are alter egos of the same person.

In addition to providing an explanation of Saul's antisubstitution intuitions, my account will, I hope, also make them seem more plausible to those inclined to reject them.

---

[5] "Substitution and Simple Sentences," p. 104.

[6] Robert M. Adams brought this point to my attention, in conversation.

II PERSONAE AND THEIR INHABITANTS

One initially attractive move, which Saul does not consider, is to attribute the substitution failures to the fact that 'Superman' and 'Clark Kent' are names of *fictional characters* Because they are fictional characters, sentences like (1), (2), and (3) are not simple after all, but contain (implicit) *fiction operators* And since fiction operators are intensional, it is no surprise that these sentences can exhibit substitution failures

But fictionality has nothing to do with substitution failure in these cases There are similar examples involving names of real people Consider the case made famous in the book and film *The Three Faces of Eve* A woman is purported to have three distinct personalities, and goes by three different names, 'Eve White', 'Eve Black', and 'Jane', depending on which personality is manifesting itself Eve White, for example, is timid and repressed, whereas Eve Black is bold and seductive Accordingly, (7) would seem to be true

(7) Eve White is more timid than Eve Black

Yet, substituting 'Eve White' for 'Eve Black' yields (7′), which is false

(7′) Eve White is more timid than Eve White

Another example, less well known (though perhaps less controversial), is that of Andre Charles, a relatively unassuming man who frequently affects the look and attitude of a glamorous female supermodel named 'Ru Paul' Thus, though Andre Charles goes by both 'Andre Charles' and 'Ru Paul', (8) seems, intuitively, to be true

(8) Ru Paul is more glamorous than Andre Charles

though, of course, (8′) is false

(8′) Ru Paul is more glamorous than Ru Paul[7]

'Eve White', 'Eve Black', 'Andre Charles', and 'Ru Paul' are not names of fictional characters, yet they exhibit the same failures of intersubstitutability, so it seems unlikely that the fact that 'Superman' and 'Clark Kent' are fictional names will figure in an adequate solution of Saul's puzzle In what follows, therefore, I shall ignore the differences between fictional and nonfictional cases

---

[7] Other nonfictional examples include the Australian actor Barrry Humphries, otherwise known as the meretricious dowager Dame Edna Everage, and the country musician Garth Brooks, who released a recording as the mysterious "international pop/rock superstar" Chris Gaines

Note that it is not simply in virtue of there being two distinct *names* that substitutivity fails, as shown by the following examples

(9) Mark Twain was a more successful writer than Samuel Clemens
(9′) Mark Twain was a more successful writer than Mark Twain

which are *both* false

There are suggestive differences among these examples The name 'Mark Twain' is not associated with the sort of distinct *behavioral profile* one associates with the names 'Eve Black' and 'Ru Paul' 'Mark Twain' is simply a *nom de plume*; whereas 'Eve White' and 'Eve Black' are associated with distinct *personalities* Ru Paul, in contrast, seems not to be a distinct personality from Andre Charles, there is not the sort of *involuntary compartmentalization* one finds in cases like Eve White/Eve Black Likewise with Clark Kent and Superman Still, the behavioral (and sartorial) characteristics one associates with the names 'Clark Kent' and 'Ru Paul' are certainly different from those one associates with 'Andre Charles' and 'Superman' The "assumed identities" of Superman and Andre Charles are more like *personae*—something along the lines of a *role* played by an actor (who will (usually) dress and behave differently when inhabiting the role), which may be voluntarily taken on and abandoned

But there are important differences between actor's roles and personae in the sense that is relevant here Compare the role of Hamlet and the persona of Ru Paul If Derek Jacobi walks by in costume backstage during a production of *Hamlet*, it seems false to say 'There goes Hamlet' what goes there is Derek Jacobi, dressed to play (and perhaps behaving like) Hamlet Whereas if Andre Charles walks by in full Ru Paul regalia, acting like a supermodel, one may well remark, truly, 'There goes Ru Paul'

So perhaps the thing to say is that Ru Paul and Clark Kent are personae—some sort of behavioral/sartorial *syndromes*; something *like* a part one plays—intermittently inhabited by Andre Charles and Superman But this will not do, for reasons analogous to the reasons Saul gives for the failure of the Fregean account At least as I have characterized them, it is not a *persona* that works at the *Daily Planet* and has a crush on Lois Personae do not do any of the things that Ru Paul and Clark Kent do, they do not *do* anything at all So if we are to take the referents of 'Clark Kent' and 'Ru Paul' to be distinct from the referents of 'Superman' and 'Andre Charles', they cannot be personae

I think this is on the right track, however, for surely personae have something to do with distinguishing Superman from Clark Kent and Ru Paul from Andre Charles Consider the case of Batman Bruce

Wayne decides that he is going to wear a leather bat costume and prowl Gotham City at night looking for criminals. He says to himself "When I am out there fighting crime, wearing this costume and using all these bat gadgets, I am *Batman*." That is, he *dubs* himself, *qua inhabitant of a particular persona he has constructed*, 'Batman'. Similarly, when Andre Charles discovered that he likes to dress as a woman and act like a supermodel, he created the persona of Ru Paul, and dubbed himself, qua inhabitant of that persona, 'Ru Paul'. Thus, 'Batman' and 'Ru Paul' refer to Bruce Wayne and Andre Charles when they are inhabiting their respective relevant personae.

Let us call a person *qua inhabitant of a persona* an *alter ego* of that person, and the person him or herself the *primum ego* of that alter ego.[8] The thesis I want to develop is that Ru Paul, Batman, et alia are alter egos, and 'Ru Paul', 'Batman', et alia are their names.

The case of Superman and Clark Kent is different in an important way. Neither Superman nor Clark Kent is, I claim, an alter ego of the other. Rather, they are *both* alter egos of the Kryptonian Kal El. Although Kal El was named 'Clark Kent' by his adoptive parents, in Metropolis he has two distinct personae, the Milquetoast reporter and the Man of Steel, and goes by 'Clark Kent' when inhabiting the former and 'Superman' when inhabiting the latter. So, one and the same person can inhabit distinct personae—can have different alter egos—at different times (Kal El is sometimes Clark Kent, sometimes Superman), or intermittently inhabit a single persona (Andre Charles is not always Ru Paul).

The recent Hollywood film *The Mask of Zorro* brings out another interesting point, for it shows that *different* people can inhabit the same persona at different times. The story has Don Diego de la Vega *passing on* the Zorro persona to Alejandro Murieta—someone else who will dress in a certain way and do certain sorts of things, and call himself 'Zorro' while he is doing them, someone, that is, who will inhabit the Zorro persona. So Zorro, once Don Diego de la Vega's alter ego, becomes Alejandro Murieta's alter ego. One can easily imagine Bruce Wayne doing the same thing with the Batman persona—handing it down to Dick Grayson, for example, or even Barbara Gordon (the police commissioner's daughter, whose alter ego is Batgirl[9]).

---

[8] 'Alter ego' and 'primum ego' are correlative terms: one is not a primum ego unless one has an alter ego.

[9] One need not *be* male in order to have Batman as one's alter ego. All that is required is that one *present oneself as* male.

Kal El, Andre Charles, Bruce Wayne, and Don Diego inhabit their personae in something *like* the way an actor plays a dramatic role, though Ru Paul and (within their respective fictions), Clark Kent, Superman, Batman, and Zorro are not fictional[10] Nor are they *characters* portrayed by their creators, like Flip Wilson's Geraldine and Paul Rubens's Pee-Wee Herman[11] When Bruce Wayne inhabits the Batman persona, he is not *playing* or *portraying* Batman, he is *being* Batman, likewise when Kal El inhabits the Superman or Clark Kent persona, and when Andre Charles inhabits the Ru Paul persona Derek Jacobi, in contrast, is not being Hamlet, he is playing him No one could *be* a fictional character

III PERSONAE, ALTER EGOS, AND THE REFERENCE OF ALTER-EGO NAMES

The suggestion, so far, is that names like 'Superman', 'Ru Paul', 'Batman', and 'Zorro' refer to alter egos, where an alter ego is an individual qua inhabitant of a persona, and an individual qua inhabitant of a persona is distinct from that individual simpliciter[12] More can be said about what sort of thing alter egos might be and how they differ from primum egos, and I shall do so in section v First, however, let us explore the notion of a persona a bit more, as well as the relation of alter-ego names to their referents

Personae, as I am construing them, consist (most often[13]) of behavioral and sartorial *guidelines*—specifications of general ways of dressing and behaving—which allow for a significant degree of freedom in their realization.[14] Thus, for a person to inhabit a persona is for that person to have, under certain conditions (see below), its characteristic properties. The properties characteristic of the Clark Kent persona include presenting oneself as male, being cowardly and feckless, wearing nondescript clothing (for example, black horn-rimmed glasses, a plain grey suit and fedora), working at the *Daily Planet,* concealing one's super powers,[15] concealing one's relation to Superman and Kal El, and so on. Among those characteristic of the Superman persona are presenting oneself as male, being courageous and capable, wearing a flashy costume (for example, a blue leotard with a big red 'S' on it and a red cape), *not* concealing one's super powers, but using them to fight crime, concealing one's relation to Clark Kent, and so on. Kal El inhabits the Clark Kent and Superman personae when he dresses and behaves in the appropriate ways. Kal El qua inhabitant of the Clark Kent persona *is* Clark Kent, and Kal El qua inhabitant of the Superman persona *is* Superman.

---

[10] That is, Clark Kent and Superman are *no more* fictional than Kal El, and Batman is *no more* fictional than Bruce Wayne (Of course, not all dramatic characters are fictional, as witness the play *Tru*, which includes Truman Capote among its *dramatis personae*)

[11] Such characters are subtly different from dramatic roles Wilson and Rubens portray Geraldine and Pee-Wee Herman as part of a performance, just as Jacobi might portray Hamlet, and (or so it seems to me) Geraldine and Pee-Wee Herman are both fictional But they are characters Wilson and Rubens created only for themselves to portray Indeed, they involve specific behavioral profiles, distinctive voices and movements that perhaps only their creators are capable of Dramatic roles, in contrast, may be performed by anyone (Though authors of dramatic roles might have rights to their characters in other ways Tom Stoppard's *Rosenkrantz and Guildenstern are Dead* could conceivably have been the subject of a lawsuit had it been written while Shakespeare was alive)

[12] And, further, an individual (a person) is different (though not intrinsically) from a primum ego as noted above, only individuals with alter egos have primum egos, a person with an alter ego is a primum ego, a person without an alter ego is simply a person

[13] One can imagine a *secret* persona, consisting of certain psychological properties—say, certain ways of thinking—that are never manifested publicly

Within limits (to be discussed below), Kal El could modify either persona and still retain the Superman and Clark Kent alter egos. Superman could, for example, sing Wagner while he flies, or wear a costume with different colors or components, without ceasing to be Superman. And Clark Kent could be less cowardly, or wear fashionable glasses or suits, without ceasing to be Clark Kent. Likewise, Ru Paul could become a soprano or stop wearing wigs without ceasing to be Ru Paul, and Batman could act more like a bat (by, say, emitting high-pitched shrieks while chasing criminals), or wear a modified bat-suit, without ceasing to be Batman.[16]

Moreover, personae are like literary creations, and alter egos are like embodied characters, such as Geraldine and Pee-Wee Herman, in that their originators have certain *rights* with respect to them, such as the exclusive right to modify them and to determine who else (if anyone) may legitimately inhabit them (and, hence, take on the corresponding alter ego). It is not the case that just *anyone* inhabiting the Superman or Clark Kent persona is Superman or Clark Kent Unless the Superman and Batman personae became public domain, no one could legitimately inhabit them or modify them without Kal El's or Bruce Wayne's permission—just as no one could legitimately portray or modify Geraldine or Pee-Wee Herman without Flip Wilson's or Paul Rubens's permission (or without the characters' becoming public domain) The *false* or *imposter* Supermans and Batmans that sometimes appear in the stories are illegitimate inhabitants of the respective personae

---

[14] This is another way in which personae differ from dramatic roles: whereas a role is *performed,* a persona is (something like) *improvised within.*

[15] Hence the truth of Saul's (12) in "Substitution and Simple Sentences," p. 105: 'Clark Kent can fly, though he conceals this fact'.

[16] Moreover, they could change the *names* of their alter egos without changing the alter egos themselves. Superman could change his name to 'Tim' without ceasing to be Superman, just as Kal El could change his name to 'Kelly' without ceasing to be Kal El.

An alter ego is, therefore, a person qua *legitimate* inhabitant of a persona, where a legitimate inhabitant is either the persona's originator or an authorized successor, and an authorized successor is either one to whom the persona is passed on by its creator (as in the case of Don Diego de la Vega and Alejandro Murieta) or who has otherwise acquired the right to inhabit it [17]

If Don Diego stops inhabiting the Zorro persona, Zorro no longer exists (except timelessly) But if he passes the rights to it on to Alejandro Murieta, and Murieta inhabits it, then Zorro exists again, though as Murieta's alter ego and not Don Diego's And if Murieta in turn passes it on to someone else, then Zorro becomes the alter ego of yet another person A persona may be inhabited by a series of individuals authorized to do so, in something like the way actors can take over roles in a series of plays, television shows, or movies This may happen either if the persona has become public domain and a new person lays claim to it, or if the persona's originator explicitly transfers the right to inhabit it. If a persona has been legitimately inhabited by *n* people, then we might say that its corresponding alter ego is the fusion of the *n* distinct people-qua-inhabitants-of-that-persona

There are, however, definite limits to the changes that a legitimate inhabitant may make to a persona, saving identity The Superman persona *necessarily* includes the properties of presenting oneself as a man and having super powers no one without super powers or presenting himself as a woman could be Superman [18] The Ru Paul persona necessarily includes the properties of presenting oneself as a woman and as glamorous no one presenting herself as male or a fashion victim could be Ru Paul And the Batman persona necessarily includes the properties of presenting oneself in a bat-like way and as a man no one presenting himself in a pelican-like way or as a woman could be Batman

---

[17] One can imagine legal confrontations arising from illegitimate or simultaneous claims to a persona Andre Charles might sue someone for impersonating Ru Paul, and if Bruce Wayne died without passing the Batman persona on, Dick Grayson and Barbara Gordon might both claim an exclusive right to inhabit it and sue each other for (something like) trademark or copyright infringement

[18] Although it does seem possible for Kal El to stop *using* his super powers to fight crime and still be Superman Moreover, it seems that when Kal El *loses* his super powers, as sometimes happens in the stories, he *does* cease to be Superman

These restrictions might seem to commit me to a Fregean description theory of alter-ego names, but they do not There is no obstacle to taking such names to be directly referential rigid designators, in spite of the fact that their bearers must have certain properties in order to bear the names It does not follow from a name's being directly referential that there are *no* properties a thing must have in order to be its referent There are, for example, no possible worlds in which Nixon is a swarm of locusts[19], but this does not mean that not being a swarm of locusts is part of the *meaning* of 'Nixon' And there are no possible worlds in which Superman is a weakling presenting himself as a woman, though it does not follow that not being a weakling and not presenting oneself as a woman are part of the meaning of 'Superman'

Consider the following counterfactual story about 'Hesperus' and 'Phosphorus' Suppose the ancient astronomers used 'the evening star' to fix the reference of 'Hesperus' and 'the morning star' to fix the reference of 'Phosphorus' They believed that Hesperus and Phosphorus are distinct heavenly bodies Now, on the actual story, when it was discovered that Hesperus and Phosphorus are the same planet, astronomers came to accept 'Hesperus is Phosphorus' as true But they might not have They might have decided to retain 'Hesperus' as the name for Venus *appearing in the evening sky* and 'Phosphorus' as the name for Venus *appearing in the morning sky* That is, they might have retained the names of Venus given under different conditions *as* names of Venus under those conditions [20] Note that it still would not be the case either that 'Hesperus' means *the evening star* or that 'Phosphorus' means *the morning star*—or, for that matter, that they would mean *Venus appearing in the evening sky* and *Venus appearing*

---

[19] Cf Saul Kripke, *Naming and Necessity* (Cambridge Harvard, 1980), p 46

[20] Cf Alan Gibbard, "Contingent Identity," *Journal of Philosophical Logic*, IV (1975) 187-221 "In rare cases, at least, one thing will be of two different kinds, with different persistence criteria, and whereas one proper name refers to it as a thing of one kind, another proper name will refer to it as a thing of another kind " (As, for example, '*Goliath*' and 'Lumpl') I need not follow Gibbard in claiming that rigidity is sortal relative, however, since in the case I am imagining 'Hesperus = Phosphorus' is *false*

in the morning sky.[21] Such descriptions would merely serve to fix the references of 'Hesperus' and 'Phosphorus'.

On this counterfactual story, the properties expressed by the reference-fixing descriptions would nonetheless be *necessary* properties of Hesperus and Phosphorus: nothing could be Hesperus that was not Venus-appearing-in-the-evening, and nothing could be Phosphorus that was not Venus-appearing-in-the-morning. This is, of course, not true in general of the properties expressed by reference-fixing descriptions (contingent properties of an object may be used to fix it as the referent of a name). But in this case it is, given that, in general, there are properties that named things must have in order to bear their names, and, in the particular case imagined here, those properties are also the properties used to fix the reference of the names.

There is an analogous counterfactual story for 'Superman' and 'Clark Kent'. Let us say that Kal El creates the Superman and Clark Kent personae for himself, but does not name them. Moreover, as he goes about the business of being a superhero he does not speak to anyone, so no one has occasion to learn his name. Now, let us say the people of Metropolis dub the guy who flies around fighting crime 'Superman', and Kal El's neighbors dub the shy guy they see walking home at night 'Clark Kent'. No one believes that Superman and Clark Kent have anything to do with each other. Now suppose that everyone finds out that it is one and the same guy, Kal El, inhabiting both of these personae. What will they do? One thing they *might* do is keep using 'Clark Kent' to refer to Kal El qua inhabitant of the Clark Kent persona and 'Superman' to refer to Kal El qua inhabitant of the Superman persona. In the actual case, Kal El dubs *himself* 'Superman' qua inhabitant of the Superman persona and 'Clark Kent' qua inhabitant of the Clark Kent persona. But I cannot see that that makes any difference in what the names name.

So, in the case of an alter-ego name, there are particular discriminating properties an object must have in order to be its bearer, though these properties are not part of the sense of the name. What is baptized at the creation of an alter ego is a person *qua inhabitant of a persona*. The properties constituting the persona at the time of the dubbing function to fix the reference of the alter-ego name, as well as to determine what properties the person must have in order to be the alter ego So an alter-ego name refers to a particular person with the relevant persona properties, persons without the persona properties are not named by the alter-ego name

---

[21] Moreover, it would not be the case that anything satisfying the descriptions 'appearing (in Venus's position) in the morning sky' and 'appearing (in Venus's position) in the evening sky' would be (respectively) Hesperus and Phosphorus. It is *that planet*, Venus, and no other heavenly body, which is Hesperus as it appears in the evening sky and Phosphorus as it appears in the morning sky.

If you are inclined to doubt the claim that an object-qua-instantiator-of-a-particular-property can be named, consider the following First, note that there are in general words—namely *predicates*—that apply to objects only qua instantiators of particular properties A predicate *F* (correctly and literally) applies only to objects instantiating *F*ness, and if *F*ness and *G*ness are distinct properties, then *F* and *G* can apply to different things, and it is false that *F*s are *G*s Take 'water' and 'snow', for example Although it is true that all snow is water, it is not true that all water is snow water and snow are not the same thing.[22] Next, note that we may construct rigid referring expressions out of these predicates—for example, 'dthat water' and 'dthat snow'[23]—that will refer a particular collection of H_2O molecules under different conditions—namely, exactly under those conditions under which the predicates differentially apply 'dthat water' designates the collection of H_2O molecules in all worlds at all times, whereas 'dthat snow' designates it only in certain worlds at certain times—that is, the ones at which 'snow' applies to it

But if these rigid referring expressions can differentially refer to one and the same collection of H_2O molecules under different conditions, then why not names? Imagine a particular quantity of snow in the shape of a man Let us call this particular snowman 'Fred' And let us say that Fred melts, leaving a puddle of water on the floor Is this puddle Fred? Well, this particular quantity of water—let us call it 'Cal'[24]—is not a snowman (it is not even snow) Puddles of water are not snowmen; so Cal is not Fred Cal and Fred do consist of the same *water*, however So we can say that 'Fred' is a name that applies to a particular quantity of water—namely, Cal—only insofar as Cal is in a particular state 'Fred', that is, applies to Cal-qua-instantiator-of-the-property-of-being-snow-in-the-shape-of-a-man

---

[22] Other relevant examples are 'stoat'/'ermine' and 'salmon'/'smolt' The stoat is a brown short-tailed weasel that is called an "ermine" *when its fur turns white during the winter* Thus, though ermines are stoats, stoats are not ermines A smolt is a young salmon at the point in its development *when it assumes the color of an adult and is ready to migrate to the sea* Although smolts are salmon, salmon are not smolts

[23] Of course, 'snow' and 'water' may themselves be rigid, if they are natural-kind terms

[24] It might be pragmatically odd to give a puddle of water a name puddles are not the sorts of things one would ordinarily *need* a name for But I cannot see that there is anything wrong with it semantically One can name anything

IV THE IDENTITY SENTENCES

On the account developed here, names of alter egos and names of their primum egos are not coreferential An important consequence of this claim is that (5), (6), and (10) come out false·

(5) Superman is Clark Kent
(6) Bruce Wayne is Batman
(10) Andre Charles is Ru Paul

For Saul, it is a condition of adequacy on any account of the substitutivity failures in sentences like (1), (2), and (3) that it preserve the truth of the identity sentences, and this is an intuition that no doubt many will share Nevertheless, I think these sentences are not true, and that this explains why the substitutions do not go through

The approach I prefer here is a pragmatic one, on which *literal* tokens of sentences such as (5), (6), and (10) (that is, tokens having the meaning of their types) are false, but *nonliteral* tokens of such sentences (that is, tokens having meanings distinct from those of their types) may be true[25]

Compare (5), (6), and (10) to (11), which appeared in a recent *New York Times* advertisement.

(11) Cathy Rigby is Peter Pan

It seems clear that what is meant by this (printed) utterance is not that Cathy Rigby *is* Peter Pan (no one could *be* Peter Pan), but that Cathy Rigby is *playing* Peter Pan in a production of the musical And we may suppose that this is what was meant without commitment either to relative identity (which would not help anyway Is the same *what?*) or to the 'is' of (11) being an 'is' of *role portrayal* (or its converse)

The critic's rave,

(12) Jacobi *is* Hamlet!

provides further support for this approach Construed literally, (12) is, like (11), absurd (no person could be a fictional character) What the utterance means is that Jacobi's portrayal of Hamlet is perfect (or very close to it), or something like that The emphatic use of a literally absurd sentence (note the stress on 'is'), in this case, has the effect of communicating the critic's astonishment at the quality of the performance It is a bit of hyper-hyperbole, like the once common graffiti,

---

[25] There are other possibilities One might, following Peter Geach—*Reference and Generality* (Ithaca Cornell, 1962), and *Logic Matters* (Berkeley California UP, 1968), claim that identity is sortal relative, and argue that on any appropriate expansion of 'is' (5), (6), and (10) come out false Alternatively, following David Wiggins—*Sameness and Substance* (Cambridge Harvard, 1980)—one might claim that identity sentences are ambiguous, and argue that (5), (6), and (10) are either identity claims and false, or true but not identity claims I prefer the pragmatic account since it avoids the controversies over sortal-relative identity and the proliferation of senses of 'is'

(13) Clapton is God[26]

Again, one need neither go relative (which would not help here, either) nor introduce a *portrays to perfection* or *ought to be worshiped as* (or whatever) sense of 'is'

On the pragmatic approach, utterances of (6) and (10) may be taken to mean the same as (6′) and (10′)

(6′) Bruce Wayne is the person whose alter ego is Batman
(10′) Andre Charles is the person whose alter ego is Ru Paul

(construed as literal identity claims) And utterances of (5) may be interpreted as meaning the same as (5′)

(5′) The person whose alter ego is Superman is the person whose alter ego is Clark Kent

(again, construed as a literal identity claim)

It is instructive to compare (14) and (15)

(14) Dr Jekyll is Mr Hyde
(15) Eve White is Eve Black

both of which are simply false Dr Jekyll is not Mr Hyde, and Eve White is not Eve Black Nor is Dr Jekyll the person whose alter ego is Mr Hyde 'Dr Jekyll' and 'Mr Hyde' are, arguably, names of distinct *persons* manifesting themselves in the same body Likewise, Eve White is not the person whose alter ego is Eve Black 'Eve White' and 'Eve Black' are (arguably) names of distinct *personalities* manifesting themselves in the same person If we take these names to be names of, respectively, distinct persons and distinct personalities, we would have a straightfoward explanation of substitution failures in cases like (7)-(7′) and (16)-(16′)

(7) Eve White is more timid than Eve Black
(7′) Eve White is more timid than Eve White
(16) Mr Hyde is meaner than Dr Jekyll
(16′) Mr Hyde is meaner than Mr Hyde

---

[26] Clapton being of course the guitarist Eric Clapton

In any case, the availability of reconstruals of (5), (6), and (10), each of which is motivated by the need to explain independent facts about the use of identity sentences, ought to weaken the case for their non-negotiability Thus, the implied falsity of (5), (6), and (10) is not *eo ipso* grounds for rejecting the account I have offered here

V WHAT ALTER EGOS MIGHT BE

I have argued that an alter ego is something like an embodied character, namely, a person-qua-inhabitant-of-a-persona But what could such a thing *be*, such that it is a different *object* from a person simpliciter?

The approach I shall pursue here is to identify alter egos with *fusions of time-slices* of the individuals whose alter egos they are [27] (Of course, Saul considers and rejects an account in terms of time-slices, since it entails that (5) and (6) are false But I have argued that this objection is not well motivated)

Note that another way of describing the relation between Kal El, Superman, and Clark Kent is to say that Kal El is Superman *when and only when* he is inhabiting the Superman persona, and is Clark Kent *when and only when* he is inhabiting the Clark Kent persona The corresponding claim about the names is that 'Superman' refers to Kal El when and only when he is inhabiting the Superman persona and 'Clark Kent' refers to Kal El when and only when he is inhabiting the Clark Kent persona And another way to supply distinct referents for the names is to take Superman and Clark Kent to be distinct fusions of distinct time-slices of Kal El

Let us call the temporal intervals during which Kal El is inhabiting the Superman persona *Superman time-slices*, and the temporal intervals during which he is inhabiting the Clark Kent persona *Clark Kent*

---

[27] Again, there are other possibilities One might simply take alter egos to be *primitive* they are just what they are, and are not reducible to anything else (This might appeal to those who are prepared to recognize the irreducible existence of fictional characters) Or one might recognize an object qua instantiator of a particular property as an entity distinct from that object qua instantiator of some other property, and from that object *tout court* Such "qua-objects"—which might also, on analogy with time-slices, be called "property-slices"—would consist of an object and one or more of its instantiated properties (perhaps construed as *tropes*) There is also the possibility of construing alter egos as *stages*, which will be discussed below As well, one might adopt an *endurantist* ontology and, along the lines set out in Trenton Merricks's "Endurance and Indiscernability"—this JOURNAL, XCI, 4 (April 1994) 165-84—construe 'Superman' as coreferential with 'Kal El at *t*', for all times *t* such that Kal El is inhabiting the Superman persona at *t* The *perdurantist* account I develop here should be taken as an example of how an ontology for alter egos might be worked out Any ontology on which they are distinct from the persons whose alter egos they are will support the solution to Saul's puzzle advocated here

time-slices. We can then identify Superman with the fusion of the Superman time-slices and Clark Kent with the fusion of the Clark Kent time-slices.[28] Since those fusions are distinct, the referents of 'Superman' and 'Clark Kent' are distinct, and the possibility of, for example, (1)'s being true is explained.

A person's alter ego is the fusion of the time-slices of the person during which she inhabits a persona—that is, during which she instantiates the properties constituting the persona. What of the person herself—the primum ego? There seem to be two options. One might construe a primum ego either as the fusion of the discontinuous segments *complementary to* the alter ego or as the fusion of the alter ego and its complements (that is, the complete space-time worm of which the alter-ego slices are slices).

The former option has objectionable consequences. Consider Bruce Wayne and Batman again. Were we to identify Batman with the fusion of the Batman-slices and Bruce Wayne with its complement, it would follow that Bruce Wayne *ceases to exist* during those times when Batman is around. But this does not seem plausible. For it is, after all, *Bruce Wayne* who is wearing the mask and cape. Moreover, on this account Bruce Wayne would go out of and into existence *at will*, which also does not seem plausible. That one's *creations* may go out of and into existence at one's will seems plausible enough (think of Penelope weaving, unraveling, and reweaving her tapestry, or Gibbard squashing *Goliath*). Batman might cease to exist when Bruce Wayne is out of persona, and would cease to exist permanently were Bruce Wayne to give up being Batman entirely with no one else to take the job. But it does not seem at all plausible that Bruce Wayne might go out of *and into* existence at *his own* will.

So let us say that an alter ego is a fusion of time-slices of an individual who continues to exist simultaneously with it. Alter egos are therefore *proper spatiotemporal parts of* the persons whose alter egos they are. And since they are proper parts of their primum egos, they are distinct from them. 'Kal El', 'Bruce Wayne', 'Andre Charles', et alia refer to complete spatiotemporal continuants, whereas 'Superman' and 'Clark Kent', 'Batman', 'Ru Paul', et alia refer to distinct fusions of distinct proper parts of them.

---

[28] Fusions rather than sets or sequences, since sets and sequences are not concrete objects, and nonconcrete objects cannot leap tall buildings or enter phone booths, whereas fusions of concrete objects, and their parts, can. Notice also that it would not be correct to identify *personae* with fusions of time-slices, since personae (or the rights to them) may be passed from one person to another, creating a new alter ego, whereas time-slice fusions cannot.

This proposal captures the intuitions about the cases quite well On the time-slice account, Superman and Clark Kent are proper parts of Kal El, but neither is a part of the other, they are distinct objects[29] Thus, (17)[30]

(17) He hit Clark Kent once, but he never hit Superman

could be true Superman never gets hit when Clark Kent does, since they are never in the same place at the same time

Moreover, since on this account what is true of an alter ego is true of its primum ego, though not vice versa, the following sorts of cases are also accommodated Since it is Bruce Wayne who wears the bat costume and drives the Batmobile, (18) is false

(18) Batman drives the Batmobile more often than Bruce Wayne

as are (19)

(19) Superman leaps tall buildings more often than Kal El

(20)

(20) He hit Superman once, but he never hit Kal El

and (21)

(21) He saw Batman once, but he never saw Bruce Wayne

Kal El leaps whenever Superman does, and he gets hit as often as Clark Kent, and Bruce Wayne is seen whenever Batman is In contrast, Kal El and Clark Kent do lots of things Superman does not do Likewise Bruce Wayne and Batman

The time-slice account also provides a principled answer to Saul's question "Which is he [Superman] when he is in the shower" (*ibid*, p 104)? The skeptical force of this question derives in part from a failure to appreciate that Superman and Clark Kent are both alter egos when Kal El is in the shower, assuming he is not dressed as or behaving as either Superman or Clark Kent, he is *neither* Superman nor Clark Kent he is just Kal El, the guy from Krypton Compare the answer to the corresponding question with respect to Batman when Bruce Wayne is in the shower, neither dressed nor behaving as Batman, he is just Bruce Wayne

---

[29] Kal El makes sure that the personae are inhabited in such a way that the alter egos are distinct This does not appear to be necessary for alter egos in general, however Properly constructed, they *might* overlap (imagine Andre Charles taking over the Batman role, but inhabiting it as Ru Paul)

[30] "Substitution and Simple Sentences," p 103

The example of Zorro is again instructive. After Don Diego passes the "rights" to the Zorro persona to Murieta, whenever Murieta inhabits it *he* is Zorro. Likewise anyone to whom Murieta might pass on the Zorro persona. So we should take alter-ego names to name fusions of persona-instantiating time-slices of possibly more than one person. The alter ego is then extended to include time-slices of its successive primum egos.

Note, further, that the time-slice account can be applied to the counterfactual Hesperus/Phosphorus case discussed above, as well as to the case of Fred and Cal. On the counterfactual scenario, Hesperus is the fusion of the evening time-slices of Venus (Venus during those temporal intervals in which appears in the evening), and Phosphorus is the fusion of the morning time-slices of Venus (Venus during those temporal intervals in which it appears in the morning). Hesperus, Phosphorus, and Venus are distinct entities, though the former are both proper spatiotemporal parts of the latter. Likewise, we can think of Fred as the fusion of time-slices of Cal during which he is snow in the form of a man.

A similar sort of story might be told about Jekyll/Hyde and Eve White/Eve Black/Jane, as well. We could identify Dr. Jekyll and Mr. Hyde with distinct fusions of time-slices of Jekyll/Hyde's *body*, and Eve Black and Eve White as distinct fusions of time-slices of Jane (the *person* who emerges during the course of treatment). Persons, personalities, and alter egos would thus be three different kinds of time-slice fusions, where the kinds are determined by the sort of conditions that individuate the relevant slices.

VI. SOME OBJECTIONS

One prima facie problem for my account is this. If Superman and Kal El are distinct objects, then it would seem that there are *two* individuals leaping tall buildings whenever Superman does, *two* individuals having the thoughts Superman has, and so on—since, after all, it is Kal El who wears the cape, and the like. Another worry is that, if alter egos are, as I have claimed, not *persons*, how could they do all the things that persons do, and how could they be physically and psychologically continuous with persons?

These concerns are easily answered. Although Kal El and Superman are distinct four-dimensional objects, they have four-dimensional parts that *overlap* (likewise Kal El and Clark Kent): any time-slice of Superman or Clark Kent is also a time-slice (of the very same spatiotemporal dimensions) of Kal El (though not, of course, vice versa). When Kal El is flying around as Superman, or reporting to work as Clark Kent, there is only one individual present, with only one set of thoughts.[31] (Note, however, that this individual is *neither* Kal El *nor* Superman *nor* Clark Kent—each of whom is temporally too large to be identical to it—but, rather, a shared temporal part of each.) Thus, though alter egos are not persons, they are identical to proper temporal parts of persons, and may do all the things that those parts may do. And for the same reason, they are psychologically continuous with persons.[32]

An intuitively satisfying way of describing the relations among Kal El, Superman, and Clark Kent is to say that, *while* Kal El is inhabiting the Superman persona, he and Superman are *identical*, while he is inhabiting the Clark Kent persona, he and Clark Kent are identical, and while he is inhabiting neither persona, he is identical to neither. (Since the personae are never inhabited simultaneously, Superman and Clark Kent are never identical.) Of course, strictly speaking, Kal El is *never* identical to either Superman or Clark Kent, since (on my account) they are distinct four-dimensional objects: what *are* identical are specific spatiotemporal proper parts of each of them. If, however, we allow that the names 'Kal El', 'Superman', and 'Clark Kent' may be used in a loose sort of way to refer to time-slices of their respective referents, then we could accommodate this way of describing the relation between Kal El, Superman, and Clark Kent.[33] We could also say that Kal El flying around as Superman (likewise Kal El reporting to work as Clark Kent) is a single individual bearing two names. And we could allow that it is true in a loose sort of way that, for example, Superman is flying over Metropolis between 12:00 noon and 1:00 p.m. on Friday, even though strictly speaking he is temporally too big to fit into that interval (what is strictly true is that a temporal part of Superman is flying over Metropolis, and so on). (This would also provide another way of taking particular utterances of the identity sentences to be true.[34])

---

[31] The appeal to overlapping temporal parts is a standard four-dimensionalist way of avoiding coincident entities. See, for example, Mark Heller, *The Ontology of Physical Objects* (New York: Cambridge, 1990), pp. 14-16.

[32] Another way to respond to these worries would be to follow Ted Sider—"All the World's a Stage," *Australasian Journal of Philosophy*, LXXIV (1996): 433-53—in *identifying* persons with person stages (that is, time-slices), and alter egos with alter-ego stages—in which case Superman (Clark Kent) and Kal El are the same person (person stage) when Kal El inhabits the Superman (Clark Kent) persona, but not when the persona is uninhabited.

[33] Cf. Heller, pp. 12-14.

[34] On a *three-dimensionalist* account of alter egos, this way of speaking could be taken to be literally true if identity were relativized to time. (The notion of time-relativized identity is explored by George Myro in "Time and Identity," in Richard E. Grandy and Richard Warner, eds., *Philosophical Grounds of Rationality*

Another possible problem for my account concerns apparent violations of Leibniz's law. If at time *t* Bruce Wayne is being Batman, and it is denied that Bruce Wayne and Batman are at *t* distinct but coincident entities, then, given Leibniz's law, anything true at *t* of Bruce Wayne ought to be true at *t* of Batman. But if there is some subsequent time *t′* at which Bruce Wayne is *not* being Batman, then there may well be things true at *t* of Bruce Wayne that are not true at *t* of Batman. For example, if on Monday Bruce Wayne is being Batman, then, if it is true of Bruce Wayne that he will have lunch with Vicki Vale on Tuesday, it ought also to be true of Batman that he will have lunch with Vicki Vale on Tuesday. Yet, it might *not* be true of Batman that he will have lunch with Vicki Vale on Tuesday—if, for example, Bruce Wayne is planning not to be Batman on Tuesday. Thus, though on Monday Bruce Wayne is Batman, Bruce Wayne and Batman do not share all of their properties.

But the violation of Leibniz's law is only apparent, since, strictly speaking, Bruce Wayne and Batman are *not* identical. If, again, we allow the loose sort of talk on which utterances of names can refer to proper parts of their bearers, then it might be true that on Monday 'Bruce Wayne' and 'Batman' pick out the same object—a shared temporal part. But in this case it would not be true on Monday of Bruce Wayne that he will have lunch with Vicki Vale on Tuesday, since the Bruce Wayne of Tuesday (a particular twenty-four-hour chunk of a four-dimensional object) is not identical to the Bruce Wayne of Monday (a distinct twenty-four-hour chunk of the same four-dimensional object). The utterance on Monday of 'Bruce Wayne' in an utterance of 'Bruce Wayne will have lunch with Vicki Vale tomorrow', if it is to corefer with an utterance of 'Batman', will have to be taken to refer to a particular temporal part of Bruce Wayne. But *that* part will not exist on Tuesday, so the utterance will be false. It is *another* part of Bruce Wayne—one that is, crucially, not also a part of Batman—that will have lunch with Vicki Vale on Tuesday. So there is no violation of Leibniz's law.[35]

Intentions, Categories and Ends (New York: Oxford, 1986), pp. 383-409. Myro credits the basic idea to H. P. Grice.)

[35] Another way to go here is, again, to follow Sider (*op. cit.*) and identify alter egos with alter-ego stages. To say that Batman will have lunch with Vicki Vale on Tuesday would, on this view, be to say that the Monday alter-ego stage denoted by 'Batman' (and 'Bruce Wayne') is related in the way appropriate to alter egos (namely, the way appropriate to persons, *qua inhabitants of personae*) to a Tuesday alter-ego stage bearing the name 'Batman' which has lunch with Vicki Vale. But this claim would be false if there *were* no Tuesday alter-ego stage having lunch with Vicki Vale related

Another prima facie problem for the time-slice account concerns *modal* statements, such as (22).

(22) Superman might have not have built the Fortress of Solitude.

For, given that Superman is a particular proper part of a particular space-time worm, (23)

(23) Superman does not build the Fortress of Solitude

the truth of which in some other possible world is sufficient for the truth of (22) in the actual world, would be true of a *different* particular part of a *different* particular space-time worm in some other possible world. But that space-time worm would not, on the account I have given, be Superman. Such problems have obvious solutions in counterpart theory.[36] But that theory is itself controversial. Here I shall simply leave this as a cost of the time-slice account of alter egos.

---

in the proper way to the Monday alter-ego stage, which would be the case if Bruce Wayne were not inhabiting the Batman persona at lunch on Tuesday.

An endurantist account of alter egos could handle these sorts of cases by following Myro's suggestion (*op. cit.*, pp. 392-93) that Leibniz's law, which, given the temporal relativity of identity, is itself subject to temporal qualification, namely,

(LL_t) $(x)(y)(t)[x=y \text{ at } t \supset (Fx \text{ at } t \equiv Fy \text{ at } t)]$

be further restricted so as to exclude "temporal properties," such as having lunch with Vicki Vale *on Tuesday*. That is, predicates substituted for *F* in the schema (LL_t) may not themselves include any temporal references. So, though it is true that Bruce Wayne will have lunch with Vicki Vale on Tuesday and false that Batman will have lunch with Vicki Vale on Tuesday, 'will have lunch with Vicki Vale on Tuesday' is not a permissible substituend for *F* in (LL_t), and (temporalized) Leibniz's law is not violated. Sentences such as (iii)

(iii) Bruce Wayne will have lunch with Vicki Vale on Tuesday

should be construed not as predications of temporal properties, but as temporally indexed predications of nontemporal properties, as in (iii′)

(iii′) On Tuesday, Bruce Wayne has lunch with Vicki Vale.

Thus, to say that it is true on Monday of Bruce Wayne that he will have lunch with Vicki Vale on Tuesday is to say that it is true on Monday that it is true on Tuesday that Bruce Wayne has lunch with Vicki Vale—or, in other words, that it is true on Tuesday that Bruce Wayne has lunch with Vicki Vale (the temporal operator 'on Monday' is in this case vacuous, since if it is true on Tuesday that *p*, then it is true on any day that it is true on Tuesday that *p*). Note that (iv) and (iii′) do not imply (v)

(iv) On Monday, Bruce Wayne is Batman.
(iii′) On Tuesday, Bruce Wayne has lunch with Vicki Vale.
(v) On Tuesday, Batman has lunch with Vicki Vale.

though of course (iv′) and (iii′) do.

(iv′) On Tuesday, Bruce Wayne is Batman.

The apparent violation of Leibniz's law involves an illicit shift in temporal operators.

[36] See David Lewis, "Counterpart Theory and Quantified Modal Logic," this JOURNAL, LXV, 5 (March 7, 1968): 113-26, and "Counterparts of Persons and Their Bodies," this JOURNAL, LXVIII, 7 (April 8, 1971): 203-11—both reprinted in his

VII. CONCLUSION

I have argued that Saul's puzzle disappears once we recognize that 'Superman', 'Clark Kent', 'Batman', 'Ru Paul', 'Zorro', and the like are names of alter egos, which are distinct from the persons whose alter egos they are. Since the names of persons and the names of their alter egos are not coreferential, there is no puzzle about why they are not intersubstitutable, even in simple sentences. I have also reviewed several proposals for the ontology of alter egos and the interpretation of identity sentences with alter-ego names as constituents, as well as their costs. As a final consideration, note that since a person and his alter ego(s) are distinct entities, the Russellian propositions expressed by (24), (25), and (26) are distinct

(24) Superman can fly.
(25) Clark Kent can fly.
(26) Kal El can fly.

and there is a straightforward explanation of how Lois can believe one and not the others which is consistent with the construal of names as directly referential. Of course, this does not help the direct-reference theorist in general, since there still remain examples like (9) and (9′) embedded in belief contexts. But it does provide a unified solution to Saul's puzzle and the corresponding puzzles about belief.

DAVID PITT

Brooklyn College/City University of New York

Philosophical Papers, Volume I (New York: Oxford, 1983); see also *On the Plurality of Worlds* (New York: Blackwell, 1986).

Note that on the time-slice account, alter-ego names would not be rigid designators, though they may still be directly referential *quasi*-rigid designators (see Lewis, *On The Plurality of Worlds*, p. 256).

NEW BOOKS ANTHOLOGIES

ASHMAN, KEITH M , and PHILIP S BARINGER, eds *After the Science Wars* New York Routledge, vii, 224 p Cloth $85 00, paper $25 95

BAGHRAMIAN, MARIA, and ATTRACTA INGRAM, eds *Pluralism The Philosophy and Politics of Diversity* New York Routledge, 2000 xiii, 267 p Cloth $85 00, paper $25 95

BEINER, RONALD, and WAYNE NORMAN, eds *Canadian Political Philosophy Contemporary Reflections* New York Oxford, 2001 vi, 408 p Cloth $65 00, paper $35 00

BENSON, PETER, ed *The Theory of Contract Law New Essays* New York Cambridge, 2001 xiv, 349 p Cloth $64 95

BOGHOSSIAN, PAUL, and CHRISTOPHER PEACOCKE, eds *New Essays on the A Priori* New York Oxford, 2000 xi, 478 p Cloth $70 00, paper $24 95

BOXHILL, BERNARD, ed *Race and Racism* New York Oxford, 2001 vi, 486 p Paper $24 95

CAPPS, DONALD, ed *Freud and Freudians on Religion A Reader* New Haven Yale, 2001 vii, 357 p Paper $18 00

CAZEAUX, CLIVE, ed *The Continental Aesthetics Reader* New York Routledge, 2000 xvii, 618 p Cloth $90 00, paper $29 99

COLTHEART, MAX, and MARTIN DAVIES, eds *Pathologies of Belief* Malden, MA Blackwell, 2000 vii, 224 p Paper $26 95

CRAIG, WILLIAM LANE, and J P MORELAND, eds *Naturalism A Critical Analysis* New York Routledge, 2000 xv, 286 p Cloth $90 00

CRANE, TIM, and SARAH PATTERSON, eds *History of the Mind-Body Problem* New York Routledge, 2001 viii, 255 p Cloth $75 00

DANCY, JONATHAN, ed *Normativity* Malden, MA Blackwell, 2000 xv, 143 p Paper $22 95

DEPAUL, MICHAEL, ed *Resurrecting Old-Fashioned Foundationalism.* Lanham, MD Rowman and Littlefield, 2001 xiii, 93 p Cloth $57 50, paper $17 95

EKSTROM, LAURA WADDELL, ed *Agency and Responsibility Essays on the Metaphysics of Freedom.* Boulder Westview, 2001 vi, 264 p Paper $25 00

FETZER, JAMES H , ed *Science, Explanation, and Rationality Aspects of the Philosophy of Carl G Hempel.* New York Oxford, 2000 xxix, 342 p Cloth $55 00

GILL, ROBIN, ed *The Cambridge Companion to Christian Ethics* New York Cambridge, 2001 xv, 290 p Cloth $59 95, paper $19 95

HARCOURT, EDWARD, ed *Morality, Reflection and Ideology* New York Oxford, 2001 ix, 193 p Cloth $35 00

HARTSHORNE, CHARLES, and EDGAR BRIGHTMAN *Hartshorne and Brightman on God, Process, and Persons The Correspondence, 1922-1945* RANDALL E AUXIER and MARK Y A DAVIES, eds Nashville Vanderbilt, 2001 vi, 184 p Cloth $39 95

HAYDEN, PATRICK, ed *The Philosophy of Human Rights* St Paul, MN Paragon, 2001 xviii, 686 p Paper $22 95

HOOKER, BRAD, and MARGARET OLIVIA LITTLE, eds *Moral Particularism* New York Oxford, 2000 xiv, 317 p Cloth $55 00, paper $19 95

LEAMAN, OLIVER, ed *Encyclopedia of Asian Philosophy* New York Routledge, 2001 xxvii, 669 p Cloth $160 00

LILLA, MARK, RONALD DWORKIN, and ROBERT SILVERS, eds *The Legacy of Isaiah Berlin* New York Review of Books, 2001 xiii, 198 p Cloth $22 95

LYNCH, MICHAEL P , ed *The Nature of Truth Classic and Contemporary Perspectives* Cambridge MIT, 2001 xvi, 802 p Cloth $75 00, paper $35 00

MACDONALD, PAUL S , ed *The Existentialist Reader An Anthology of Key Texts* New York Routledge, 2000 vi, 346 p Cloth $85 00, paper $24 95

MARTINOT, STEVE, ed *Maps and Mirrors Topologies of Art and Politics* Evanston, IL Northwestern, 2001 xxxix, 342 p Cloth $79 95, paper $29 95

MEYER, MICHEL, ed *Questioning Derrida With His Replies on Philosophy* Burlington, VT Ashgate, 2001 vii, 122 p Cloth $64 95

OYAMA, SUSAN, PAUL E GRIFFITHS, and RUSSELL D GRAY, eds *Cycles of Contingency Developmental Systems and Evolution* Cambridge MIT, 2001 xii, 377 p Cloth $50 00

PASSERIN D'ENTRÈVES, MAURIZIO, and URSULA VOGEL, eds *Public and Private Legal, Political, and Philosophical Perspectives* New York Routledge, 2000 vi, 201 p Cloth $75 00, paper $22 99

POLT, RICHARD, and GREGORY FRIED, eds *A Companion to Heidegger's* Introduction to Metaphysics New Haven, Yale, 2001 xvii, 342 p Cloth $40 00, paper $18 00

PREUS, ANTHONY, ed *Essays in Ancient Greek Philosophy VI Before Plato* Albany SUNY, 2001 250 p Paper $18 95

PREYER, GERHARD, and FRANK SIEBELT, eds *Reality and Humean Supervenience Essays on the Philosophy of David Lewis* Lanham, MD Rowman and Littlefield, 2001 xiii, 243 p Cloth $68 00, paper $27 95

ROUNER, LEROY S , ed *If I Should Die* Notre Dame University Press, 2001 xvii, 192 p Cloth $34 00

SEDLEY, DAVID, ed *Oxford Studies in Ancient Philosophy, Volume XIX* New York Oxford, 2000 372 p Paper $35 00

SMILEY, TIMOTHY, ed *Mathematics and Necessity Essays in the History of Philosophy* New York Oxford, 2001 x, 166 p Cloth $35 00

STERBA, JAMES P , ed *Social and Political Philosophy Contemporary Perspectives* New York Routledge, 2001 xiv, 481 p Cloth $90 00, paper $27 95

STEUP, MATTHIAS, ed *Knowledge, Truth and Duty Essays on Epistemic Justification, Responsibility, and Virtue* New York Oxford, 2001 ix, 256 p Cloth $49 95

STUMP, ELEONORE, and NORMAN KRETZMANN, eds *The Cambridge Companion to Augustine* New York Cambridge, 2001 xv, 307 p Cloth $59 95, paper $21 95

WARBURTON, NIGEL, JON PIKE, and DEREK MATRAVERS, eds *Reading Political Philosophy Machiavelli to Mill* New York Routledge, 2000 x, 382 p Cloth $75 00, paper $24 95

THE JOURNAL OF PHILOSOPHY
VOLUME XCVIII, NO. 11, NOVEMBER 2001

7/1/02

BAYESIAN EPISTEMOLOGY AND EPISTEMIC CONDITIONALS ON THE STATUS OF THE EXPORT-IMPORT LAWS*

The notion of probability occupies a central role in contemporary epistemology and cognitive science. Nevertheless, the classical notion of probability is hard to reconcile with the central notions postulated by the epistemological tradition.[1] Bayesian epistemologists have presented three main types of responses to this problem. First, there are *eliminative strategies*: if notions like belief cannot be probabilistically articulated, they have no theoretical use.

---

* I am grateful to Rohit Parikh, Teddy Seidenfeld, and Bas van Fraassen for helpful comments and suggestions. This work was supported in part by a grant from the National Science Foundation.

[1] A classical reference in this area is Henry Kyburg, *Probability and the Logic of Rational Belief* (Middletown: Wesleyan, 1961). One of the paradoxes presented in Kyburg's book, the so-called *lottery paradox*, is now well known in the literature and will be considered here in detail. The moral of the paradox can be summarized as follows: it shows that, if one accepts as epistemically rational to believe propositions just in case it is rational to have a high degree of confidence in them, then one has to accept as well that it is epistemically rational to be in weakly inconsistent epistemic positions, where currently held propositions do not have a nonempty intersection. Epistemologists who insist on logical closure and consistency as requirements for rational belief reject the possibility of this consequence, and, therefore, they also reject modeling qualitative doxastic states in terms of threshold value for degree-of-confidence functions. Of course, one's philosopher modus ponens can be easily transmuted into another's philosopher modus tollens. So, epistemologists, like Kyburg himself, have argued that the scenario of the lottery gives us good reason to adopt the threshold value model and to reject (the full strength of) logical closure as a doxastic ideal of rationality. The latter solution is compatible with a probabilistic stance in epistemology (where probability is the only primitive), but it delivers a notion of belief that does not connect smoothly with the received view in the epistemological tradition. The former solution is incompatible with probabilism due to the fact that it conflicts with the adoption of acceptance rules in terms of high probability, or even probability one (or probability infinitesimally close to one).

0022-362X/01/9811/555–93

and, consequently, they have to be *eliminated* from a mature epistemology Second, there are *pluralist strategies*, which recommend adopting both belief and probability as irreducible primitive notions Finally, there is the claim that a unified Bayesian epistemology can be articulated around the sole postulation of a primitive notion of conditional probability I shall focus here on two aspects of this third probabilistic strategy First, I shall define both belief and belief change from conditional probability—following an idea first proposed by Bas van Fraassen[2] Second, I shall argue that the Adams hypothesis (on how to accept conditionals[3]) can be better articulated and extended in this setting Finally, I shall discuss both the internal commitments and the range of applicability of the probabilistic view of belief and conditionals which thus arises

I INTRODUCTION

van Fraassen[4] has tried to find a common ground between probabilism and the received view in epistemology "What I hope for is some reconciliation of the diverse intuitions of Bayesians and traditionalists, within a rather liberal probabilism" (*ibid*, p 170) In a more technical article, van Fraassen both reconsidered the motivations for pursuing this form of *unified probabilism* and proposed various concrete manners of formulating it

> Personal or subjective probability entered epistemology as a cure for certain perceived inadequacies in the traditional notion of belief But there are severe strains in the relationship between probability and belief They seem too intimately related to exists as separate but equal, yet if either is taken as the more basic, the other may suffer [ ] I would like to propose a single unified account, which takes conditional personal probability as basic (*op cit*, p 349)

Epistemologically, the idea is to appeal to a third notion (besides qualitative belief and subjective grading), namely, the notion of *supposition,* and to encode it via a pre-Kolmogorovian notion of conditional probability

> There is a third aspect of opinion, besides belief and subjective grading, namely *supposition* Much of our opinion can be elicited only by asking us to suppose something, which we may or may not believe Here

---

[2] "Fine-grained Opinion, Probability, and the Logic of Full Belief," *Journal of Philosophical Logic,* XXIV (1995) 345-77

[3] Formulated by Ernest Adams in "Probability and the Logic of Conditionals," in Jaakko Hintikka and Patrick Suppes, eds , *Aspects of Inductive Logic* (Amsterdam North Holland, 1966), p 273

[4] *Laws and Symmetry* (New York Oxford, 1989)

> supposition will be a central ingredient of an unified probabilistic picture (*op cit* , p 351)

One of my main goals here is to articulate, extend, and explore the limits of this unified form of probabilism I shall consider not only how to define qualitative belief and expectations in a paradox-free manner, but also how these doxastic notions change when new items are supposed or learned I shall characterize the correspondent notion of *probability-based belief change*, and compare it with nonprobabilistic accounts of belief change recently developed in the literature

It is important to realize that the *noneliminative probabilism* just considered has been postulated not only in order to define a unified epistemological view, but also in order to *reform and improve the traditional notion of probability itself* The following example can help to clarify this point

> Consider a map of the Western Hemisphere and focus your attention on any two points on the map Say that you select a point N representing New York City and that you choose a second point B representing Baltimore In addition, call W the surface representing the Western Hemisphere, S the surface representing the state of New York, and M the surface representing Maryland Now pick a point P at random in W The (unconditional) probability that P is either N or B is zero, since P has been chosen from an infinite number of points In spite of this fact, it seems that one would refuse to say that the probability that the arbitrarily chosen point is N is undefined, given that the point is either N or B Moreover, it seems rather natural to say that such probability is exactly 1/2[5]

Two main strategies have been implemented to solve this problem[6] On the one hand, one can take *infinitesimal numbers* at face value One can say that the proposition that the randomly chosen point P is N carries an *infinitesimal probability* ι Such probability is a nonstandard number greater than zero but strictly smaller than any standard positive number By the same token, the probability that the point is either N or B carries $2i$ value And the conditional probability that P is N given that P is either N or B would then carry the value $i/2i$ This representation of the problem begins to approximate intuition Now, the controversial probability of P being N *given that* P is either N or B *is* defined In addition, infinitesimal numbers obey arithmetical laws that allow us to cancel the infinitesimal i in the quotient $i/2i$ And this yields the desired result 1/2

---

[5] This example is inspired by a similar example initially proposed by A N Kolmogoroff, and by a variant of it recently proposed by Adams in his *A Primer of Probability Logic* (Stanford CSLI, 1988), pp 252-53

[6] There are, to be sure, alternative approaches to the ones I am about to present, but they are not available to epistemologists of probabilistic persuasion I shall focus here on solutions that can be formulated in the context of a probabilistic epistemology Section II, on the other hand, begins to articulate the distinctions that separate relevant varieties of Bayesian epistemology With the help of these distinctions, I shall also sketch the gist of alternative approaches to the ones considered in this section

At first sight, the move might appear as a rather natural response to the problem of how to condition on events of (standard) measure zero Nevertheless, the strategy causes as many problems as it seems to solve Many philosophers and mathematicians have for centuries considered infinitesimal numbers with suspicion As a result, infinitesimals have been often replaced in applications by limits of series of standard numbers At least this was so until Abraham Robinson made them respectable again in 1966 with his important *Non-standard Analysis* [7] Today, the theory of infinitesimal probability is well understood, and many philosophers have officially adopted it as an improvement of standard Bayesian theory, for example, Brian Skyrms, David Lewis, and Vann McGee[8] have recently argued in favor of this view

Several well-known properties of ordered algebraic fields are violated in this new setting A salient one is the so-called *Archimedean axiom* establishing that for any positive number r, no matter how small, there is an integer n such that $nr > 1$ A non-Archimedean field will contain at least an infinitesimal element ε with the property that $n\varepsilon < 1$ for every positive integer n, even though ε is itself positive A nice property of non-Archimedean subjective probability is that it obeys all the Kolmogorovian axioms extended to nonstandard values Nevertheless, many foundational problems remain open, for example, infinitesimals are not measurable

The most salient alternative to the use of infinitesimal numbers is based upon the work of Karl Popper and Alfred Renyi [9] In standard probability theory, absolute probability is adopted as a primitive, and characterized axiomatically via Kolmogoroff's axioms Conditional probability is then *defined* in terms of absolute probability Both Popper and Renyi proposed to study a theory where this idea is completely inverted[10] Conditional probability is adopted as a primitive, and absolute probability is defined as conditional probability given the truth[11] Then the notion of conditional probability is axiomatically characterized in such a way that the probability of a proposition *A* given *B* can be meaningfully defined even when the unconditional probability of *B* is zero van Fraassen appealed to a variant of the Popper-Renyi strategy in order to construct a unified version of probabilism

---

[7] Amsterdam North-Holland, 1966

[8] Skyrms, *Causal Necessity* (New Haven Yale, 1980), p 175, Lewis, "A Subjectivist's Guide to Objective Chance," in William Harper, Robert Stalnaker, and Glenn Pearce, eds , *Ifs* (Boston Reidel, 1980), pp 267-99, McGee, "Learning the Impossible," in Ellery Eells and Brian Skyrms, eds , *Probability and Conditionals Belief Revision and Rational Decision* (New York Cambridge, 1994), pp 179-99

[9] Popper, *The Logic of Scientific Discovery* (New York Harper, 1959, second edition, 1965), new appendices *ii to *v, Renyi, *Foundations of Probability* (San Francisco Holden-Day, 1970)

The Popper-Renyi view has its own foundational problems, some of which are considered in sections VIII and IX below Nevertheless, the Popper-Renyi's *two-place probability functions* can be interestingly applied when the underlying probabilistic space is at most of countable size In addition, a result recently proved by McGee (*op cit*) shows that the standard values of infinitesimal probability functions are representable as *two-place probability functions*, and that every two-place function is representable in terms of the standard real values of some infinitesimal probability function McGee's arguments show that there is a robust and hidden unity linking the two strategies used to define *generalized conditional probability* Both strategies capture aspects of the intuitive notion of probability, which the post-Kolmogorovian account seemed to have failed to capture The trade-off is that the new strategies lack solid foundations in measure theory My goal here is not to close this foundational gap, but to examine the claims that have been made in favor of the use of two-place functions As we just saw, one important step toward the formulation of a unified probabilism is to produce a paradox-free account of belief and belief change. Another important challenge for the formulation of a unified probabilism is to produce an adequate account of how we accept and use conditional sentences.

---

[10] The underlying motivations of Popper and Renyi were different, in addition, the theories they proposed differ in important aspects that I cannot review here in detail

[11] The underlying idea here is that there are no degrees of belief defined in a vacuum All belief is conditional given some body of background knowledge "Unconditional" degree of belief is, therefore, an artificial limit case where the judgments are made against no background of *empirical* hypotheses But, of course, the probabilistic "status quo" includes all a priori logical truths This is a consequence of the fact that probabilistic representations of attitudes are affected by the so-called problem of *logical omniscience* All probabilistic agents are perfect logicians, or, to put it in a different manner, the minimal admissible background knowledge a probabilistic agent can admit is one where he is certain of all logical laws Philosophers have different views as to the reasonability of this feature of probabilistic representations I want to remark that logical omniscience will not be put into question here It should be said in passing also that nothing precludes saying that the probability of *B* given *A* is the probability that *B* is true, when *A* is a tautology—see D Edgington, "Lowe on Conditional Probability," *Mind*, CIV, 420 (October 1996) 617-30

Frank Ramsey[12] articulated the view, now common, according to which the conditional probability of *B* given *A* should be viewed as the degree of belief of *B* under the supposition that *A* (*ibid.*, p. 76). He was also responsible for a very influential view of conditionals, which he understood as "cognitive carriers."[13] According to Ramsey's view, conditional sentences lack truth conditions, but they have precise acceptability conditions, which can be formulated inside a unified probabilistic epistemology. The idea is that to evaluate a conditional we add its antecedent hypothetically to our stock of knowledge, and then assess the consequent on that basis. As Ramsey puts it, in that case we "are fixing our degree of belief in [the consequent of the evaluated conditional] given [the antecedent]." But, "if [the antecedent] turns out false, these degrees of belief are rendered *void*" (*ibid.*, p. 154, footnote). We can appeal to the geographical example provided above in order to illustrate this point. What is the probability of the conditional, 'If the randomly chosen point *P* coincides either with *N* or *B*, then it lies either on the surface representing New York (*S*) or on the surface representing Maryland (*M*)'? Ernest Adams,[14] in his own formulation of Ramsey's account, has proposed to set this probability to one by default; in this case, the default value seems to match intuition. Nevertheless, consider the following two conditionals: (1) 'If the randomly chosen point *P* coincides either with *N* or *B*, then it lies on the surface *S*', (2) 'If the randomly chosen point *P* coincides either with *N* or *B*, then it does not coincide with *N*, and it does not coincide with *B*'.[15]

One might argue that the probability of both conditionals is defined, and that the probability of the first conditional is 1/2 and the probability of the second is zero rather than one McGee has recently proposed the same remedy as the one considered above in order to treat this problem, that is, he proposed an improved version of the Ramsey-Adams hypothesis (degree of acceptability of conditionals go by conditional probability) where the underlying notion of conditional probability is generalized conditional probability Here I intend to extend and articulate further McGee's approach An important limitation of his proposal is that it does not cover iterated conditionals, so I shall provide models of acceptance for those conditionals I shall also claim that this extension provides valuable insights about the nature of conditional inference induced by unified probabilism (and on the nature of unified probabilism itself) The analysis of the so-called export-import axiom ('If ϕ, then if ψ, then χ' if and only if 'If ϕ and ψ, then χ'[16]), for example, will shed light on the unique features of the probabilistic construal of conditional inference In fact, although I shall argue that probabilistic models of conditionals require the law, almost no rival account endorses it[17] This applies not only to the most salient accounts in terms of possible worlds (for example, the models offered by Lewis, Robert Stalnaker, Donald Nute, or John Pollock[18]), but also to the so-called *epistemic* models in terms of qualitative notions of supposition[19] In addition, several articles questioning the validity of the export law for English conditionals have been recently published in this JOURNAL[20] I shall show, nevertheless, that export-import is a deeply entrenched commitment of unified probabilism

---

[12] "Truth and Probability," in *Philosophical Papers* (New York: Cambridge, 1990), pp. 52-94.

[13] This view is clearly expressed by Ramsey in "General Propositions and Causality," in *Philosophical Papers*, pp. 145-63.

[14] See "Probability and the Logic of Conditionals," p. 273.

[15] Adams's classical account of conditionals appears in *The Logic of Conditionals* (Boston: Reidel, 1975). Nevertheless, in this book Adams avoids making a decision about what to do with probability zero antecedents. He uses assignments where every factual sentence has positive probability and where the language does not contain conditionals with inconsistent antecedents.

It is perhaps important to see what *type* of argument I am offering Rather than defend the a priori validity of the law or the fact that

---

[16] The variable χ ranges over conditional sentences eventually containing conditionals, while ϕ and ψ range over conditional-free sentences

[17] Probabilists have offered arguments in favor of the law (usually by appealing to qualitative models) McGee, for example, offers an argument of this type in "A Counterexample to Modus Ponens," this JOURNAL, LXXXII, 9 (September 1985 462-71 Frank Jackson also defended the law in his *Conditionals* (New York Blackwell, 1987), appendix A1, p 133 Edgington also argued in favor of exportation and importation in her state of the art article on conditionals in *Mind*, CIV, 414 (1995) 235-329 Adams has a more detached view, prima facie favorable to the law

[18] Lewis, *Counterfactuals* (Cambridge Harvard, 1973), Stalnaker, "A Theory of Conditionals," in Frank Jackson, ed , *Conditionals* (New York Oxford, 1991), pp 28-45, Nute, *Topics in Conditional Logic* (Boston Reidel, 1980), Pollock, *Subjunctive Reasoning* (Boston Reidel, 1976)

[19] An exhaustive review of both theories can be found in C Cross and Nute, "Conditional Logic," in D Gabbay and F Guenthner, eds , *Handbook of Philosophical Logic, II Extensions of Classical Logic* (Boston Kluwer, 2001, second edition)

[20] The most recent is Bernard Katz, "On a Supposed Counterexample to Modus Ponens," this JOURNAL, XCVI, 8 (August 1999) 404-15

English conditionals obey it, I shall focus on showing that the law arises as a natural commitment for defenders of certain forms of unified probabilism I shall also argue that the problem of finding interconnections between qualitative and quantitative doxastic states is intimately related to the problem of clarifying the account of conditionals induced by unified probabilism In particular, I shall show that the validity of the export-import law itself depends on robust properties of probability-based belief change

As van Fraassen suggests, much of our opinion (belief, belief change, and conditionals, among other notions) can be elicited from suppositions I intend here to make this process of elicitation and reduction explicit It is also my purpose to reveal the important epistemological constraints imposed by it In a nutshell, two of my main interests here are first, to show how a unified probabilism can be possible, and, second, to determine which are some of its most salient commitments

II MONISM AND PLURALISM IN BAYESIAN EPISTEMOLOGY

Richard Jeffrey[21] has recently coined the term 'radical probabilism' I hope to delimit this view from other forms of Bayesian epistemology and to make clear what the differences (similarities) are separating (uniting) these approaches There is a "family resemblance" linking van Fraassen's and Jeffrey's varieties of radical probabilism, which can be described in Jeffrey's own words

> Radical probabilism doesn't insists that probabilities be based on certainties, it can be probabilities all the way down, to the roots (*ibid*, p 11) [ ] Radical probabilism adds the 'non-foundational' thought that there is no bedrock of certainty underlying our probability judgments (*ibid*, pp 44-45)

This form of "nonfoundationalism" is shared between van Fraassen's and Jeffrey's forms of radical probabilism Unlike Bruno De Finetti or L J Savage, they appeal to a form of *probabilistic monism* according to which certainties (or full beliefs) are not used in order to determine the space over which rational agents attribute probabilities Both theories adopt a radical stance by reducing the set of admissible epistemological primitives to just a single probabilistic concept Any other epistemological notion is legitimate only if it can be derived or explicated in terms of this single primitive In contrast, De Finetti[22] has eloquently described a *pluralistic strategy* in Bayesian epistemology

---

[21] *Probability and the Art of Judgment* (New York Cambridge, 1992)

[22] *Theory of Probability, Volume I* (New York Wiley, 1990)

> In almost all circumstances, and at all times, we find ourselves in a state of uncertainty Uncertainty in every sense [ ] It would therefore seem natural that the customary modes of thinking, reasoning and deciding should hinge explicitly and systematically on the factor *uncertainty* as the conceptually pre-eminent and determinative element The opposite happens however there is no lack of expressions referring to uncertainty, but it seems that these expressions, by and large, are no more than verbal padding The solid, serious, effective and essential part of arguments, on the other hand, would be the nucleus that can be brought within the language of certainty—of what is certainly true or certainly false It is in this ambit that our faculty of reasoning is exercised, habitually, intuitively and often unconsciously (*ibid*, p 24)

Immediately, De Finetti makes clear that his set of certainties contains more than mere tautologies, and that its main role is to determine a space of possibilities

> Thinking of a subset of truths as given (knowing, for instance, that certain facts are true, certain quantities have given values, or values between certain limits, certain shapes, bodies or graphs of given phenomena enjoy certain properties, and so on), we will be able to ascertain which conclusions, among those of interests, will turn out to be—on the basis of the data—either *certain* (certainly true), or *impossible* (certainly false), or else *possible* (*ibid*, p 25)

What about probability? According to De Finetti "probability is something that can be distributed over the field of possibility"

> Using a visual image, which at a later stage could be taken as an actual representation, we could say that the logic of certainty reveals to us a space in which the range of possibilities is seen in outline, whereas the logic of the probable will fill in this blank outline by considering a mass distributed upon it (*ibid*, p 25)

De Finetti usually refers to this assumed body of certainties as the *state of information* of the agent It is important to realize that the adoption of a pluralist stance does not preclude recognizing the cognitive primacy of the notion of conditional probability

> We have at times insisted on making clear the fact that every prevision, and, in particular, every evaluation of probability, is conditional, not only on the mentality or psychology of the individual involved, at the time in question, but also, and especially on the state of information in which he finds himself at that moment (*ibid*, p 134)

Notable subscribers of this pluralistic view include I J Good, Levi, and Savage[23] Notice that full belief (certainty) and conditional probability are introduced here as two independent primitives In a sense certainty is the true primitive Probability can be defined only when the dual of certainty (possibility) is specified Moreover, in this framework it is perfectly possible to study several static and dynamic properties of bodies of certainty independently of the probabilistic aspects of the model In other words, *belief change is not determined by probability change* This is the degree of independence missing when a monist stance is adopted Even if one can specify in an unproblematic manner belief as a function of probability, belief change is now completely determined by probability change The correspondent notion was called above *probability-based* belief change Most of the work in the first part of the present discussion will be devoted to the clarification of the properties of probability-based belief change induced by *unified probabilism,* that is, by a noneliminative probabilistic strategy[24]

III ADAMS'S HYPOTHESIS AND ACCEPTABILITY OF CONDITIONALS

Adams has proposed the following hypothesis in a series of articles and books since, at least, the mid 1960s (*op cit*)[25]

> *Original Adams hypothesis* the probability of a simple conditional 'If *a*, then *b*' will be $cp(B|A) = P(A \cap B)|P(A)$, if $P(A)$ is nonzero, and 1 otherwise—'*A*' denotes the proposition expressed by '*a*', and the same applies to '*B*' and '*b*'[26]

In most versions of Bayesian epistemology, *credence* in a proposition *X* is the probability measure of the degree of partial or full belief in *X* *Conditional credence,* on the other hand, is represented in terms of the conditional probability *cp* As D H Mellor[27] points out in a recent article, conditional credences are not credences $P(A \cap B)|P(A)$ is a function of two credences, not a credence in its own right. Conditional credences are, nevertheless, probabilities (they satisfy standard axioms). So, we can adopt the following reformulation of the Adams hypothesis (due to Mellor (*ibid.*, p. 233)): the degree of *acceptance* of 'If *a*, then *b*' is determined by the conditional credence $cp(B|A)$, as long as such conditional credence is defined. As I explained above, the default clause (for the case in which $cp(B|A)$ is undefined) is unreasonable. A remedy is to start with a primitive notion of conditional probability, and to constrain it with suitable axioms, allowing for well-defined conditional probabilities $P(A|B)$, where B can carry probability zero. Following recent terminological conventions, I shall call such probabilities *two-place probability functions*. The new test is then reformulated as follows:

---

[23] See Good, *The Estimation of Probabilities An Essay on Modern Bayesian Methods* (Cambridge MIT, 1965) Levi has articulated this view in several essays A detailed explanation can be found in *The Enterprise of Knowledge* (Cambridge MIT, 1983) The standard reference to Savage is *The Foundations of Statistics* (New York Dover, 1972, second revised edition)

[24] Jeffrey's radical probablism is a paradigmatic example of an eliminative type of probabilistic epistemology

[25] See also McGee's presentation of the Adams hypothesis in "Learning the Impossible" As I explained above, the Adams hypothesis appears in its most complete form in "Probability and the Logic of Conditionals"

[26] I shall present more detailed conventions about notation at the end of section II

[27] "How to Believe a Conditional," this JOURNAL, XC, 5 (May 1993) 233-48

> *Improved Adams hypothesis*: the degree of acceptance of 'If *a*, then *b*' is given by $P(B|A)$, where $P(—|—)$ is a two-place probability function.

And, of course, one can distill a test for *full acceptance* of conditionals from this improved version of the Adams hypothesis. In fact, this test for acceptance will just be a limit case of the improved Adams hypothesis:

> *Improved Adams hypothesis for acceptance*: 'If *a*, then *b*' is accepted with respect to $P(—|—)$ if and only if $P(B|A) = 1$.

According to Adams, degrees of acceptance go by degrees of conditional credence. So, these degrees of acceptance are not credences on their own right, but they are probabilities. According to the improved version, degrees of acceptance (and full acceptance) are encoded as degrees of a primitively given notion of conditional probability. This new account preserves the formal features of probability in countable spaces, but it might not be proper to refer to it as a *probability measure* either. According to the received view in conditional probability, if w is a point in a probability space with universe U, $P(—|w)$ is not probability given an (arbitrary) *event*, but probability given a *sigma field*. A more comprehensive discussion of this point is offered in sections viii and ix. Here I want to point out only that the adoption of the improved version of the Adams hypothesis requires a further departure from the idea that the measure of acceptance of conditionals is a probability measure. The original Adams hypothesis preserves the idea that the measure of degrees of acceptance of conditionals, even if not a credence, is a probability measure (it is a conditional credence). The adoption of the improved version removes various problems related to the original version, but it would be better to say that the degrees of acceptance it characterizes are given by a *probability function* (rather than a probability measure). Although this is an important foundational issue, it will not be of central importance until we consider infinite probability spaces.

The Adams hypothesis is sometimes presented in a slightly different manner. Mellor (*ibid.*, p. 234), for example, presents the Adams hypothesis for acceptance as follows: the acceptance of 'If *a*, then *b*' is given by $cp(B|A) \approx 1$, where '$\approx$' indicates that the value of the conditional credence is at least close to 1, when $cp(B|A)$ is defined.[28] How close the conditional credence $cp(B|A)$ is (or should be) to 1 is not an issue that matters in these formulations. It would be useful briefly to compare this view with the one just sketched above. First, the improved Adams hypothesis for acceptance does not appeal to $cp(\mathrm{B}|\mathrm{A})$, but to a primitively defined notion of conditional probability. An agent can have definite views about the acceptance conditions of 'If *a*, then *b*', even when $P(A)$ is zero.[29] The original thesis cannot deal with those cases, while the improved version can. Even when the conditional credence $cp(B|A)$ is undefined, the agent can have a view as to the degree to which B is expected upon *supposing* A. And this can be determined by $P(B|A)$. Second, the alternative formulation is equivalent to a view according to which full acceptance of 'If *a*, then *b*' requires that the conditional probability of B given A is *infinitesimally close* to one. In other words the improved acceptance test can be alternatively formulated by appealing to infinitesimal probability

---

[28] Mellor presents this version of Adams for expository and argumentative purposes. The positive view he defends differs from the one articulated in this version of the Adams hypothesis, and to some extent from the one presented here, although there are various important coincidences between Mellor's view and mine, which will be discussed below.

[29] Mellor makes an interesting point about acceptance (*op. cit.*, p. 236). The use of a primitively defined notion of conditional probability is usually invoked in order to deal with the acceptance conditions of conditionals whose antecedents have measure zero. But there are other important set of cases, namely, all future conditionals used in decision making. The evaluation by Oswald of 'If Oswald doesn't kill Kennedy, someone else will' can be an important part of Oswald's process of deciding whether to make true the proposition O that he does not kill Kennedy. For O's truth "depends on his decision, which he has not yet made and which we may consistently suppose him quite unable to predict. That is, he has and believes he has no credence for [O] high or low or even indeterminate" (*op. cit.*, p. 236). Yet Oswald can accept of 'If Oswald doesn't kill Kennedy, someone else will', and this mental act might have an impact in his process of decision making. It seems that this type of conditionals, 'If *h*, *q*', where the agent has no view as to the value of $P(H)$, cannot be straightforwardly handled by an unmodified version of unified probabilism. Possible solutions (within the boundaries of unified probabilism) will be discussed below.

> *Nonstandard probability test for acceptance* the simple conditional 'If *a*, then *b*' is accepted with respect to a nonstandard value probability function *p* if and only if std $(p(B|A)) = 1$, where std(x) is the unique standard number which differs from *x* by at most an infinitesimal amount[30]

Third, the adoption of a primitively given notion $P(—|—)$, or an infinitesimal probability function *p*, eliminates the need for the 'otherwise' clause in the formulation of the Adams hypothesis This has various beneficial consequences, which will be the object of attention in the following paragraphs There are two standards of validity in probabilistic semantics One of them is genuinely probabilistic in kind[31]

> *Logic of likelihood criterion* an inference is *probabilistically valid* if and only if, for every positive ε, there exists a positive δ such that, under any probability assignment under which each of the premises will have probability greater than $1 - \delta$, the conclusion will have probability at least $1 - \varepsilon$

There is a salient limit case by taking the limit when the uncertainties go to 0

> *Logic of certainty criterion* an inference is *strictly valid* if and only if its conclusion has probability 1 under any probability assignment under which its premises each has probability 1

The logic of likelihood criterion describes a plausible set of inferences excluding many suspect patterns validated by the material conditional (for example, the so-called "paradoxes of material implication") Nevertheless, if one appeals to the original Adams hypothesis, the strictly valid inferences are not the probabilistically valid inferences The strictly valid inferences collapse into the materially valid inferences As McGee points out "in determining that the strictly valid inferences are the classical ones, what is important is [ ] the default condition that assigns the conditional probability 1 when the conditional probability is undefined "[32] Then he shows that appealing to the improved hypothesis eliminates this anomaly In fact, in this case the inferences probabilistically validated are exactly the ones that are strictly validated An important conclusion follows Consider the following pattern of inference *I*

---

[30] In "Learning the Impossible," McGee offered a back-and-forth theorem relating nonstandard functions *p* and two-place functions *P* Thomason and I extended this result to iteration in "Iterated Probability Kinematics," *Journal of Philosophical Logic* (forthcoming) In the presence of both results, it is clear that for each *p* used in the test of a conditional via the nonstandard probability test for acceptance, there is a two-place *P* yielding identical acceptance conditions via the improved Adams hypothesis for acceptance And, obviously, for each *P* satisfying the improved Adams hypothesis for 'If *a*, *b*,' there is $NS(P)$ such that $P(B|A) = 1$ if and only if std$([NS(P)](B|A)) = 1$

[31] My terminology is the one used by McGee in "Learning the Impossible "

$$a^1, a^2, \ \ , a^n, a_1 > b_1, \ \ , a_n > b_n \, / a > b$$

where a^i, a_i, b_i, a, and b are conditional-free formulae of the propositional language L The letters A, B, C, , stand for the propositions[33] expressed by the sentences a, b, c, , of L[34] I shall also use an extension LC of L containing conditionals of the shape $a > \beta$ where a is in L and β is in LC α, β, γ, , stand for sentences of LC Notice also that for sentences of $LC \setminus L$ there is no defined corresponding propositions This reflects agnosticism about the existence of conditional propositions

McGee has shown that any inference exhibiting the shape displayed in I is probabilistically valid if and only if it is strictly valid, as long as we use the improved hypothesis[35] This is a powerful reason to argue that the improved Adams hypothesis is the most parsimonious rendering of the original Adams hypothesis I shall argue that there are additional advantages associated with this formulation In fact, I shall argue that the alignment between strict and probabilistic validity permits us to represent valid probabilistic inferences in terms of purely qualitative operations on the beliefs naturally associated with each two-place function This representation makes it possible to bridge an important gap separating probabilistic and traditional epistemology

My argument has two steps The first step is simple Strict validity, and, therefore, probabilistic validity, can be cashed out in terms of our test of acceptance In fact, in order to check strict validity of a pattern I, one need only verify that the conditional $a > b$ is *accepted* whenever each item in the antecedent of I is *accepted* And acceptance of any conditional $a_i > b_i$ goes by the *generalized* conditional probability of B_i given A_i being 1

---

[32] "Learning the Impossible," p 190

[33] Given an underlying space $S = \langle U, F \rangle$ with U a nonempty set and F a sigma field on U, propositions are constructed as subsets of 2^U

[34] In other words, the function that takes lower-case letters and yields corresponding upper-case letters is a map taking sentences of L into 2^U, in such a way that for every sentence a of L, A is the proposition expressed by a

[35] See theorem 3, page 190, of "Learning the Impossible "

The second step is to show that the acceptance of each conditional $a > b$ with respect to a two-place probability function P is expressible in terms of purely qualitative operations on the bodies of belief associated with P and $P[A]$ the result of updating P with A Therefore, I shall show that *probabilistic validity can be expressed in terms of probability-based belief change* This is the hard part of the argument Establishing this second part requires a prior understanding of the (problematic) connections between probability and belief as well as a previous grasp of the relationships between probability kinematics and belief change The following section focuses on this issue

IV TWO-PLACE FUNCTIONS, FULL BELIEF, AND PROBABILISTIC CORES

A *space* is a pair $S = \langle U,F \rangle$ with U a nonempty set and F a sigma field on U A one-place probability measure on S is characterized by Kolmogoroff's axioms

(1) $0 \leq P(A) \leq 1 = P(U)$,
(2) $P(A \cup B) - P(A \cap B) = P(A) + P(B)$ (finite additivity),
(3) If En $n = 1,2,$, are disjoint, with union E, then $P(E) = \Sigma_n P(En)$ (countable additivity)

Conditional probabilities can be defined in the usual manner Alternatively, one can take two-place probability functions as primitives This can be done as follows a two-place probability function $P(-|-)$ on space S is a map $F \times F$ into real numbers such that

I *Reduction axiom* the function $P(-|A)$ is either a probability function on S or else has a constant value 1 The latter function will be called *abnormal*
II *Multiplication axiom* $P(B \cap C|A) = P(B|A)\ P(C|B \cap A)$, for all A, B, C in F[36]

If $P(-|A)$ is a one-place probability function, A will be called normal and abnormal otherwise Let '+' mark exclusive disjunction of propositions $A + B = (A - B) \cup (B - A)$ When two-place functions are taken as primitives, one-place probability functions are defined as follows $pr(A) = P(A|U)$ Two facts are worth keeping in mind if A is normal, so are its supersets, and if A is abnormal, so are its subsets In van Fraassen's construction, the notion of the a priori is the opposite of the idea of abnormality In fact, A is *a priori for P* if and only if $P(A|X) = 1$ for all X, if and only if $U - A$ is abnormal for P Now we can introduce the notion of superiority A is *superior* to B (denoted $A >_P B$) if and only if $P(A|A + B) = 1$ Armed with this notion, we can now introduce the following crucial concept K is a *belief core* (for P) if and only if K satisfies the following axioms

---

[36] The axiomatic use of the multiplication rule has a long pedigree in the theory of probability Harold Jeffreys, for example, assumes it as a fundamental axiom of probability (under the name 'W E Johnson's *product rule*') in his *Theory of Probability* (New York, Oxford, 1961, third edition), p 25

(A1) K is normal
(A2) if A is a nonempty subset of K while B and K are disjoint, then $A >_P B$
(A3) the complement of K, $U - K$, is normal

The following property follows from the axioms and definitions (*finesse*) all nonempty subsets of K are normal In addition, van Fraassen[37] has shown that belief cores are nested The proof makes essential use of (A2) and finesse van Fraassen has also suggested the following epistemological interpretation Say that an agent's epistemic state is represented by a probability function P Say that P induces a system of cores with nonempty intersection B The proposal is to see B as the strongest proposition (fully) believed by the agent For the moment, nevertheless, we cannot define the full beliefs of an agent as the propositions entailed by the innermost core of his system of cores (if the agent has cores) In fact, even if the agent has cores, such an intersection is not obviously guaranteed to exist Nevertheless, the following lemma shows that a non-coreless function always has a nonempty intersection of cores

> *Descending chains* the chain of belief cores induced by a non-coreless two-place function P cannot contain an infinitely descending chain of cores[38]

Therefore, with the help of this lemma we can use the following crisp definition of full belief (fitting van Fraassen's epistemological interpretation of core systems) A is a *full belief* (for P) if and only if A is entailed by P's innermost core (or A is a priori) The following section will be devoted to van Fraassen's account and to slight modifications and extensions of it

---

[37] "Fine-grained Opinion, Probability, and the Logic of Full Belief," p 356

[38] See my "Qualitative and Probabilistic Models of Full Belief," in S R Buss, P Hájek ,and P Pudlák, eds , *Logic Colloquium '98 Lecture Notes in Logic XIII* (Natick, MA Peters and ASL, 2000), pp 25-43 The lemma of descending chains (lemma 2) is stated and proved in pp 36-37, together with relevant historical remarks, pointing to related work in the field

IV 1 Zero probability and epistemic possibility van Fraassen's construction has an interesting behavior in spaces whose universe is countable Let a point p in the universe U of a space S be called *heavy* for a function P if $pr(p) = P(p|U) > 0$

> *Lemma of heavy points* if a space $S = \langle U,F \rangle$ is such that U is countable and normal for a non-coreless function P defined over S, the set of heavy points of P constitute P's smallest core[39]

van Fraassen's construction intends to avoid the identification of full belief and probability 1 in various useful models Moreover, as McGee clearly explains, one of the main motivations behind the use of two-place functions is to avoid identifying epistemic impossibility and measure zero

> If a proposition is utterly impossible epistemically, we shall assign it the probability 0, but it is by no means the case that any proposition to which we assign the probability 0 will be epistemically impossible[40]

Nevertheless, as long as 'epistemically impossible' means "incompatible with the full beliefs for P," the characterization of full belief offered in the previous section yields an identification of probability zero and epistemic impossibility At least this is so if one focuses on countable probability spaces—via the lemma of heavy points Here is a concrete example giving intuitive support to the idea that the aforementioned identification is problematic (when the universe is countable)

> Example the sample space has a countable set of atoms, resulting from the following in independent trials, a fair coin is flipped until we get a head, and then the trials stop The set of possible outcomes is indexed by the number of tails $X = (0, 1, 2, \ , n, \ , \Omega)$, where Ω designates never stopping, that is flipping forever and seeing only tails Evidently $P(X$ is finite$|U) = pr(X$ is finite$) = 1$ and $pr(X = n) = 2^{-(n+1)}$ $(n = 0, 1, \ )$, so that (obviously) $pr(X = \Omega) = 0$, and this is the only null event, apart from the impossible event

Notice that, according to van Fraassen's model, there is only one core for P $U - \Omega = C$, constituted by P's set of heavy points But it seems that P should have two "cores" C and U, and it also seems that the agent's full beliefs should be represented by U rather than by C

---

[39] The result is presented by Rohit Parikh and myself as lemma 3 1 in "Two-place Probabilities, Beliefs, and Belief Revision," in Paul Dekker, ed , *Proceedings of the Twelfth Amsterdam Colloquium* (Amsterdam ILLC, 1999), pp 1-6

[40] "Learning the Impossible," p 179

The agent might be "almost certain" that the outcome is finite, but not fully certain. The space could have admitted physically impossible events (say, that the coin after being tossed stops and stands still in mid-air). Any of those (abnormal) events would have carried zero probability, like Ω. But it seems that there should be a distinction between physically impossible events and Ω. It is clear that none of the former events should be deemed as seriously possible. It is less clear that the same criterion applies to Ω. This intuition cannot be captured inside of an unmodified version of van Fraassen's system for two reasons. First, according to condition (A3) of cores, the complement of a core should be normal. Therefore, U can never be a core. Second, the definition of full belief mandates to interpret the innermost core as the set of full beliefs. In "Two-place Probabilities, Beliefs and Belief Revision" (*op. cit.*), Parikh and I show that both obstacles can be removed via a slight modification of the mathematical apparatus used by van Fraassen. Here is the reformulation we proposed: first, redefine a core as a subset K of U obeying

Normality. K is normal.
Strong superiority condition. if A is a nonempty subset of K while B is disjoint from K, $P(B|A \cup B) = 0$.

These two conditions are enough to entail finesse. Moreover, it can be shown in general (*op. cit.*, lemma 2.4) that there is always a smallest core (the intersection of cores) and a largest core (determined by the union of cores). In addition, since the requirement that the complement of a core should be normal has been abandoned, the union of cores can perfectly be the universe of the probabilistic space. Then the universe U in the example presented above (accommodating all possible trials, including the "never stopping" event Ω) is now a permissible core of the function *pr* used in the example.

A second modification can be suggested by noticing that the lemma of heavy points guarantees that the innermost core of a function P is the strongest proposition carrying probability 1 (as long as the space over which P is constructed has a countable and normal universe U, for P). The intuition resulting from the coin-tossing example is that defining full belief as what is entailed by the innermost core of a two-place function P conflicts with the aforementioned formal fact. The agent represented in the example might at most *expect* the outcome of a trial to be finite. But it seems excessive to require that she ought to believe fully that X is finite. Using a terminology that has some currency in Bayesian epistemology, one can say that the agent can be at most "almost certain" that the outcome will be finite, but not fully certain. One possible solution is to adopt the following definitions *A* is a *full belief* (for *P*) if and only if *A* is entailed by *P*'s outermost core (or *A* is a priori) *A* is *expected* for *P* if and only if *A* is entailed by *P*'s innermost core

To be sure, expectations, as defined above, are identical with probability 1 in countable spaces[41] Moreover, the notion of full belief collapses with the a priori in such spaces So, one can perfectly well analyze the countable case armed with only two epistemological notions probability 1 and the a priori The innermost core encodes the strongest probability 1 proposition (via the lemma of heavy points) What seems unintuitive in the case under consideration is to maintain simultaneously the characterization of full belief as what is entailed by the innermost core So we have abandoned such characterization in general[42] The following diagram can help us to present in a succinct manner the framework we are suggesting in the countable case

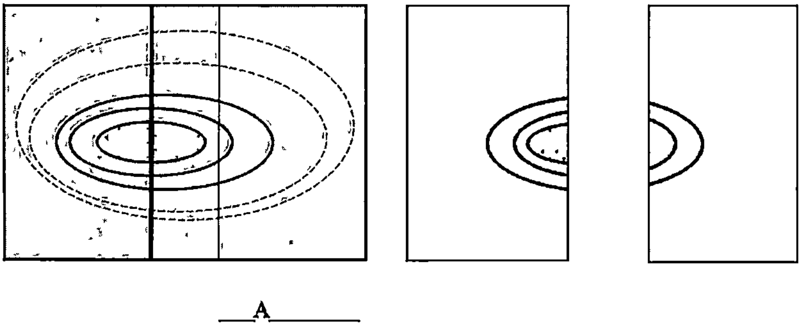


Figure 1

Focus first on the (shaded) leftmost rectangle, which represents the universe *U* of a countable space *S* The boldfaced ellipses represent the system of cores for a two-place function *Q* over *S* The largest core *F* encodes *Q*'s associated full beliefs Now the subsystem of cores embedded in this largest core can be seen as a series of *inductive*

---

[41] It is important to keep firmly in mind throughout this essay that all the introduced epistemic notions are part of the conceptual apparatus of the particular brand of radical probabilism under consideration The notion of expectation might play a different role in the context of a pluralist epistemology

[42] One of the main differences between van Fraassen's model and the model proposed here is that van Fraassen does not focus on countable models On the contrary, the countable case is of negligible interest for him His motivation is to capture the doxastic commitments of physical theories, and infinite spaces are crucial for this application The focus here is on applying unified probabilism to improve existing models of conditionals, and for this application the countable case is crucial I shall argue in sections viii and ix that Adams's ideas are hard to extend to infinite spaces

expansions[43] of Q's full beliefs ordered by strength The innermost core I can then be interpreted as the strongest possible expansion of F I is constituted by all of Q's heavy points, while all points outside I are "light" points of probability zero But Q carries enough qualitative information to distinguish among those "light" points First of all, there are probability zero events which are epistemically possible (those overlapping F), and there are probability zero events which are epistemically impossible (those not overlapping F) Of course, all fully believed events carry probability 1, but not every probability 1 event is fully believed (for example, any subcore of F carries probability 1, but is not fully believed)

It is not difficult to see that every core of a function Q carries probability 1, and, therefore, so does the largest core of Q—given that the space is countable But the central feature of the notion encoded via the largest core is that it is *robust with respect to suppositions*, in the sense that it carries probability 1 and continues to carry this probability under any supposition In fact, it is not difficult to see that $pr(F) = 1$, and $Q(F|X) = 1$, for any X This robustness with respect to suppositions is the idea captured by van Fraassen's technical term 'a priori' The weaker doxastic notion entailed by the innermost core also carries probability 1, but it is not robust with respect to suppositions For example, $Q(I|A) = 0$, where I is the innermost core in figure one and A is the rightmost cell in the partition of the leftmost diagram I propose to call the notion encoded by the innermost core *expectation* (almost certainty) [44]

A final remark is needed before concluding this section Notice that according to my account, only certain events of probability zero are *entertainable* as candidates for possible updates Those entertainable events should be epistemically possible, in the sense that they should overlap the largest core F Epistemic models in terms of a two-place probability function P give us no guidance for modeling revisions of the full beliefs for P In order to do that, we would need a gradation of the states in the complement of the largest core F This has been pictorially represented in the leftmost rectangle of figure one via several dotted ellipses The structure and epistemological function of such gradation is the target of the so-called theories of *belief revision* The reader unfamiliar with this literature can think about this ordering of the points in the complement of F as a new

---

[43] Let F and A be two compatible events in U Then $F \cap A$ is the expansion of F with A Each subcore C of F can then be seen as a expansion of F with C

[44] This terminological decision is grounded on an established notational practice in philosophical logic and computer science

"system of cores" now centered in the outermost core F This grading is usually interpreted as manner of encoding the degrees of *epistemic entrenchment* of full beliefs Such grading is quite different in nature from the one I am considering here In fact, the cores for a function P are not viewed as an entrenchment relation for full belief, but as a set of inductive expansions of the union of cores (that is, the agent's full beliefs) ordered by strength The strongest expansion is viewed as the strongest set of maximally likely propositions or expectations On the other hand, the propositions outside of the body of full beliefs are not entertainable They are not candidates for updating The system of cores for Q can, nevertheless, be interpreted in our model as yielding the epistemic entrenchment of the expectations in the innermost Q-core (if such core exists) If we see things in this way, then the nature of our modeling reveals an important feature of two-place functions, namely, the manner in which the entrenchment of expectations for a function Q is tightly connected to a set of possible expansions of the set of full beliefs for Q (the cores for Q) If we focus on the case where countable additivity is enforced, the set of "closest" A-points to the expectations of a function Q always exists as long as A is epistemically possible—even when A and I are incompatible This set can be seen as the revision of the set of expectations for Q with A, or, alternatively as the boldest inductive expansion of the body of full beliefs for $Q[A](X|Y)$, where $Q[A]$ stands for the revision of Q with the proposition A In neither case does this operation have anything to do with a "genuine" revision of the full beliefs for P[45] This matter, rather than the one I analyze here, is the main topic of the so-called theories of belief change

IV 2 On the dynamics of cores The general questions under investigation in this section are the following (1) Which is the structure of the system of cores $S[A]$ of $P[A](X|Y) = P(X|Y\cap A)$? (2) If S is the system of cores for P, which is the relationship between $S[A]$ and S? This is the last issue concerning cores that needs to be addressed before considering the probabilistic constraints on acceptance (of conditionals) induced by two-place functions

> Observation (1) if S is the system of cores of $P(—|—)$ and A is epistemically possible for P, then all the elements of $S[A] = \{C\cap A \quad C\in S$ and $C\cap A \neq \varnothing\}$ are exactly all the elements of the system of belief cores for $P[A]$[46]

---

[45] A "genuine" revision would require an input incompatible with the current corpus of full belief

[46] For a proof, see my "Hypothetical Revision and Matter-of-fact Supposition," *Journal of Applied Non-Classical Logic* (forthcoming) A web version is available via the

A picture of core change is given by the rightmost and center rectangles in figure one The shaded rectangle in this figure is partitioned by three events Let the partition cells be called *P1*, *P2*, and *P3* (where *P1* is the leftmost partition cell) The center picture represents the result of updating by *P1*, the rightmost picture the result of updating by *P3* The basic idea is that when a system of cores is updated by an epistemically possible event E the web of inductive expansions intersecting E is completely preserved

One of the many interesting consequences of observation (1) is that it can be established in a general manner without using either countable additivity or the restriction of the space to the countable case It is, therefore, a fairly robust property of two-place probability functions that the inductive structure they induce on full belief is preserved without further refinements every time they are updated

Some limit cases and new notation are needed at this point $(P[A])[B]$ will be abbreviated by $P[A, B]$ Whenever A fails to overlap the largest core of P, or when P is abnormal, we set $P[A]$ to the abnormal P assigning probability 1 of every event In addition, the innermost core of $P[A]$, denoted by $Ex(P)$, is, in this limit case, arbitrarily set to $\varnothing$ If $F(P)$ denotes P's largest core, we also set $F(P)$ to $\varnothing$ by convention

IV 3 Selecting the right dynamics Section IV 2 started with the question Which is the structure of the system of cores $S[A]$ of $P[A]$ $(X|Y) = P(X|Y \cap A)$? But one might quarrel about the relevance of the question itself In fact, there is certain freedom in defining $P[A]$ Given any P and any proposition A one can postulate (or explicitly construct) a selection function ϕ yielding the "closest" function $\phi(P, A)$ needed to change P with A[47] Positing that $P[A](X|Y) = P(X|Y \cap A)$ is tantamount to focusing on a specific manner of characterizing probability change This particular definition can perhaps be defended by its naturalness as well as by established practice in the field[48] The main reason for choosing it here will be provided in the coming section The target of section IV is to offer a representation of the improved Adams hypothesis for acceptance in terms of probability-based change I shall show that, if $P[A](X|Y) = P(X|Y \cap A)$ and

---

Computer Research Repository, Los Alamos e-Print Archive, ACM and NCSTRL (http //xxx lanl gov/abs/cs AI/0003022)

[47] This idea has been considered by Peter Gärdenfors in *Knowledge in Flux* (Cambridge MIT, 1988), chapter 5, and previously by Wolfgang Spohn in "The Representation of Popper Measures," *Topoi*, v (1986) 69-74

[48] See, for example, Richard Bradley, "More Triviality," *Journal of Philosophical Logic*, XXVIII (1999) 129-39

the underlying space is countable, then $P(B|A) = 1$ if and only if the expectations for $P[A]$ entail B

This result offers a complete qualitative characterization of the acceptance of non-nested probability conditionals for probability spaces of countable size It also clarifies the nature of the (traditional) epistemological notions tacitly involved in acceptance tests of probabilistic kind This suggests the interests of considering extensions of the model in order to study iteration where the characterization of $P[\;]$ remains unchanged Perhaps there are interesting applications of alternative models where a selection function is used in order to characterize $P[\;]$ Nevertheless, it is not clear that such models play a central role in representing qualitatively the Adams hypothesis

V SIMPLE CONDITIONALS

Before considering the iterated case, it would be beneficial to focus first on the structure of the improved Adams hypothesis for acceptance, introduced in section II

> *Improved Adams hypothesis for acceptance* a simple conditional $a > b$ is accepted with respect to $P(—|—)$ if and only if $P(B|A) = 1$

A similarly motivated idea can be expressed via an acceptance test closest in spirit to the so-called "Ramsey tests" used in recent epistemic models of conditionals [49]

> *Ramsey test for simple probability conditionals* a simple conditional $(a > b)$ is accepted with respect to $P(—|—)$ if and only if the smallest core of $P[A]$ entails B

Now the act of supposing is qualitatively represented in terms of operations on propositions In fact, the smallest core of $P[A]$ yields the expectations for $P[A]$ So, $(a > b)$ is accepted if and only if B is hypothetically expected upon supposing A [50] Notice that in this case the numerical properties of P only play a derived role They supply the structure needed in order to generate a system of cores Such system is the only structure required in order to implement the qualitative test In other words, if we were given only the system of cores for P, this would be sufficient to use the qualitative test All other representational properties of P, over and above its induced system of cores, are unnecessary in order to determine acceptance via the qualitative test The qualitative test and the Ramsey test do not coincide in general Nevertheless, they do coincide for countable probabilistic spaces

---

[49] See the section devoted to epistemic models of conditionals in Corss and Nute, see also Ramsey's "General Propositions and Causality," p 154, footnote

[50] In footnote 29 above, I mentioned an example (presented by Mellor) about the use of future conditionals used in decision making An agent might accept one of those future conditionals 'If *a*, *b*' even when he has no prior $P(A|U)$ Cases of this sort cannot be handled by an unmodified version of the improved version of the Adams hypothesis In fact, using the system of cores induced by a (total) primitive two-place function P in order to deliver acceptance conditions for 'If *a*, *b*' would require having a defined value for $P(A|U)$, even when this value is zero So, Mellor's challenge to the original Adams hypothesis seems to be also a challenge for the improved version of the hypothesis A possible solution within the boundaries of probabilism would be to represent the consistent suppositions of the agent via a

> *Coincidence lemma* if the universe U of the underlying space is countable and normal for a function P, then for all propositions A, B, $P(B|A) = 1$ if and only if the smallest core of $P[A]$ entails B[51]

It is interesting to notice that the Ramsey test for simple conditionals is perfectly well defined for infinite spaces with more than countable points In fact, the lemma of descending chains guarantees the existence of the smallest core for P and for $P[A]$, as long as A overlaps the largest core of P Nevertheless, there is no guarantee that the Ramsey test and the improved Adams hypothesis should coincide for

---

partial two-place function P—'consistent' here means that there is at least a core system delivering acceptance conditions for all the represented suppositions Or we can use a set S of (total) two-place functions sharing the defined values of the partial function P If the intersection B of the propositions representing full beliefs for each total two-place function in S is nonempty, one can also infer that such proposition B represents the full beliefs for the *partial* two-place function P

[51] Here is a sketch of the proof of the coincidence lemma Assume $Ex(P[A]) \subseteq B$ Then if $F(P) \cap A \neq \varnothing$ we have that $Ex(P[A])$ is nonempty (by the lemma of descending chains) and $P[A](Ex(P[A])|U) = 1$ Therefore, the assumption yields $P(Ex(P[A])|A) = 1 = P(B|A)$ When $F(P) \cap A = \varnothing$, $P[A]$ is abnormal Therefore, $P(X|A) = 1$ for all X In particular $P(B|A) = 1$ Assume now $P(B|A) = 1$ Then $Ex(P)$ is the support $S(P) = \{x \in U \; P(x) > 0\}$—by the lemma of heavy points So, if $A \cap F(P) \neq \varnothing$ we have $Ex(P[A]) = S(P[A])$—by observation (1) Assume then by contradiction that $Ex(P[A])$ is not included in B The proof can then be completed by cases First, consider the case $Ex(P[A]) \cap B$ is empty In this case, for all y in B, $P[A](\{y\}|U) = 0$—again by heavy points Now, since the space is countable and normal for P, countable additivity guarantees that $0 = \Sigma P[A] \; (\{y\}|U) = P[A](B|U) = P(B|A) = 1$ Contradiction Second, consider the case when $Ex(P[A]) \cap B$ is nonempty In this case, there is a strict subset Z of $Ex(P[A])$, such that $1 = P(B|A) = P(Z|A) < 1$ Contradiction Third, in order to finish the proof, we should consider the case $A \cap F(P)$ is empty This case is immediate, because by definition $Ex(P[A]) = \varnothing \subseteq B$

spaces other than the ones we are studying here (countable spaces) I shall suggest (in section viii) that the Ramsey test might be better able to preserve the intuitions of the Adams hypothesis for acceptance in those spaces Concerning the countable case, the offered analysis supports the idea that simple probability conditionals should be understood as *expectation conditionals* of the shape 'If *a*, then it is expected that *b*' [52] Different variants of this idea have been defended in recent papers and monographs [53]

I warned the reader in the introduction that, when considered in general, the qualitative properties of belief change induced by two-place functions are at odds with some of the established standards in the field This is more apparent once iteration is allowed Nevertheless, a first discrepancy from the standard models appears at the noniterated level In fact, these models are not *consistently preserving* in general This property requires that the update of a consistent view with a consistent input should generate a consistent output But when the model is not universal (that is, when the largest core does not coincide with U), updates with nonentertainable propositions yield the abnormal function This nonclassical feature will be more evident by studying iterated models This will be done in section vi

VI ITERATION AND NEGATED CONDITIONALS

My immediate concern here will be to make precise what a probabilistic model is The first component of the model is a probabilistic space $S= \langle U,F\rangle$ I shall consider tuples of the form $M = \langle S, C, \langle Ex, F\rangle, Sup\rangle$, where S is a space, C is a set of non-coreless two-place probability functions defined over S obeying the axioms I and II,[54] and *Sup* is a function from C to T, the theories definable over LC The pair $\langle Ex, F\rangle$ contains the function *Ex*, which, when applied to any P in C, yields the smallest core of P (intuitively, P's expectations) and the function F, which, when applied to any P in C, yields the largest core of P (intuitively, P's full beliefs) Finally, [ ] is a function mapping probability functions on probability functions We do not require the closure of C under [ ] A notational remark so far I have used the letters A, B, C, , in order to denote the propositions expressed by the letters a, b, c, , of L In addition, let LC be the smallest language such that $L \subseteq LC$ and if $a \in L$, $\beta \in LC$, then $a > \beta \in LC$ and $\neg\beta \in LC$ From now on we shall assume that A_M is the proposition expressed by a in model M I shall also use the constants *true* and *false*, and assume that, for every model M, $true_M = U$

---

[52] This is so, at least, for spaces whose universe is normal for the relevant function P constructed over them The important special case when the largest core coincides with the universe U of the space is studied in "Hypothetical Revision and Matter-of-fact Supposition " The resulting model for (one shot) change of the innermost expectations is, in this setting, well known It obeys the so-called Alchourrón-Gardenfors-Makinson axioms, standardly assumed in the field—see Gärdenfors Parikh and myself have also shown (*op cit*) that the conditionals validated by the Ramsey test satisfy the axioms of the theory of *nonmonotonic* reasoning presented by D Lehmann and M Magidor in "Rational Logics and Their Models A Study in Cumulative Logics," Technical Report TR 88-16, Department of Computer Science, Hebrew University of Jerusalem (November 1988)

[53] Gärdenfors and David Makinson, "Nonmonotonic Inference Based on Expectations," *Artificial Intelligence*, lxv, 2 (January 28, 1994) A similar idea appears also in several papers by Levi, as well as in his *For the Sake of the Argument Ramsey Test Conditionals, Inductive Inference, and Nonmonotonic Reasoning* (New York Cambridge, 1996) See also Andre Fuhrmann and Levi, "Undercutting and the Ramsey Test for Conditionals," *Synthese*, ci, 2 (1994) 157-69 Nevertheless, the similarities between these proposals and the one suggested above should not be overstressed Detailed comparisons (which are beyond the scope of this article) can be found in "Two-place Probabilities, Beliefs, and Belief Revision "

Various restrictions on those models have been proposed First, there are the *countable* models, where the universe U of the space S contains countable many elements Second, there are the *universal* models where, for each P in C, the largest core for P is U Third, there are the *normal* models, where all points in U are normal for any P in C In addition, all models obey

Ramsey test for probability conditionals $a > \beta \in Sup(P)$ if and only if (1) $b = \beta$ is in L and the smallest core for $P[A]$ entails B, or (2) β is in $LC\backslash L$ and $\beta \in Sup(P[A](X|Y)) = Sup(P(X|Y \cap A))$

Ramsey Test for negated probability conditionals $\neg(a > \beta) \in Sup(P)$ if and only if (1) $b = \beta$ is in L and the smallest core for $P[A]$ does not entail B, or (2) β is in LC and $\beta \notin Sup(P[A](X|Y)) = Sup(P(X|Y \cap A))$

The basic idea of the test for negated conditionals is that the acceptance of $\neg(a > \beta)$ reveals that the agent thinks that β is not accepted in the minimal hypothetical change needed to suppose consistently that A is the case In probabilistic terms when $\beta = b$, the idea is that $\neg(a > \beta)$ is accepted with respect to P whenever $P(B|A) \neq 1$ For *decisive* probability functions our proposal is tantamount to establishing that $\neg(a > b)$ is accepted with respect to P whenever $P(B|A) = 0$ But, in my framework, this is just a limit case[55]

---

[54] And where the derived one-place functions obey axioms 1, 2, and 3

[55] Notice that the model precludes being in suspense toward conditionals This epistemic feature can be relaxed by ceasing to require the test for negated conditionals

A conditional $a > \beta$ is *strictly valid in a model* $M = \langle S, C, \langle Ex, F\rangle, Sup\rangle$, if and only if $a > \beta \in Sup(P)$ for all probability functions P in C A conditional $a > \beta$ is *strictly valid* if it is valid in all models Finally, a conditional $a > \beta$ is *strictly valid in a class of models E* if it is valid in all *E*-models Strict validity can be extended to arbitrary formulas of *LC* a sentence $\alpha \in LC$ is *strictly valid in a model* $M = \langle S, C, \langle Ex, F\rangle, Sup\rangle$ if and only if $\alpha \in Sup(P)$ for all probability functions P in C[56]

Intuitively, $Sup(P)$ yields the set of conditionals accepted (or supported) by each probability function P in C together with all the logical consequences of such set The interpretation of nonconditional sentences of L, contained in $Sup(P)$, is an entirely different matter Ultimately, the propositions expressed by these sentences are propositions *accepted* by the represented agent Various forms of radical probabilism will, of course, represent acceptance of propositions in terms of high probability In the present setting, this will require stipulating that proposition X is accepted with respect to P if and only if X is entailed by the smallest core of P (or, in other terms, if X is expected) It is not surprising perhaps that this is exactly Adams's position in this matter, given that this view is compatible with the original Adams hypothesis In fact, Adams has recently suggested that we treat equally the information carried by A and by $true > a$ In other words, he suggested requiring $a \leftrightarrow true > a$[57] This can be captured in our theory by imposing the following additional postulates (1) If $a \in L$ and $a \in Sup(P)$, then $Ex(P) \subseteq A$, (2) $Th(Ex(P)) \subseteq Sup(P)$—where '$Th(Ex(P))$' yields the L-theory corresponding to the proposition $Ex(P)$ These two additional constraints essentially tell us that all nonconditional information supported by a two-place function P is constituted by the expectations for P

---

[56] Evaluating the probability of mixed formulae like $a \rightarrow (true > a)$ is a complicated matter, both in the case in which conditional propositions are admitted and when they are not The present proposal circumvents the problem by appealing to the syntactic closure of support sets So, checking the strict validity of $a \rightarrow (true > a)$ requires verifying whether the formula belongs to all possible support sets It is noteworthy, however, that even when the offered model provides acceptance conditions for mixed formulas and embedded conditionals, it does not assume that *these formulas* have probabilities In this respect it resembles acceptance models for conditionals in terms of high probability—see, for example, Adams's model in "On the Logic of High Probability," *Journal of Philosophical Logic*, xv (1986) 25579 So, for example, if α and β are simple conditionals, $(\alpha \vee \beta) \notin Sup(P)$, indicated that neither conditional is accepted with infinitesimally high probability, and where the sentence $(\alpha \vee \beta)$ is not assumed to carry probabilities In addition, the offered model articulates the doxastic commitments involved in acceptance tests in terms of high probability Acceptance tests for simple conditionals can characterize *degrees of acceptance* More complex models can only deploy the attitudes of acceptance and rejection

[57] See *A Primer of Probability Logic*, p 154

It is clear that $(\mathit{true} > a) \rightarrow a$ is strictly validated in the presence of constraint two—it is assumed here that $a \in L$ Nevertheless, the converse is not strictly valid Any function P whose associated expectations contain both A and $\neg A$ points constitutes a counterexample Nevertheless, $a \rightarrow (\mathit{true} > a)$ *is negatively validated* in our models as long as constraint one is imposed A sentence $\alpha \in LC$ is *negatively valid in a model* $M = \langle S, C, \langle Ex, F\rangle, Sup\rangle$ if and only if $\neg\alpha \notin Sup(P)$ for any probability function P in C That the notions of negative and strict validity come apart in probabilistic models of the type we are considering should be clear from the status of $a \rightarrow (\mathit{true} > a)$ [58]

Now, of course, characterizing acceptance of propositions in terms of "almost certainty" is not an appealing option for unified probabilism In fact, the view of acceptance in terms of "almost certainty" is exactly the naive view of acceptance that this new form of radical probabilism tries to reform, given that it is affected by lottery paradoxes It is more reasonable to characterize acceptance in terms of full belief, that is, in terms of what is entailed by the largest core of a two-place probability function Adopting this point of view is compatible with the endorsement of the following two constraints, instead of constraints (1-2) (1f) if $a \in L$ and $a \in Sup(P)$, then $F(P) \subseteq A$, (2f) $Th(F(P)) \subseteq Sup(P)$

If we adopt this second option, the relationships between $(\mathit{true} > a)$ and a become more tenuous Only $a \rightarrow (\mathit{true} > a)$ remains negatively valid But the converse does not hold, due to the fact that expected items need not be fully believed It seems that this second option is the one compatible with the endorsement of the "improved Adams" and with the tenets of unified probabilism

VI 1 Probabilistic and epistemic models The model presented in section VI applies to embedded conditionals and to negations of probability conditionals Many probabilists have defended the validity of inferential patterns of this sort, even when their arguments tend to appeal to possible-worlds models or to other nonprobabilistic machinery The present model also articulates probabilistic acceptance in terms of purely qualitative operations on propositions, but those operations are probabilistically justified via the connection between qualitative and quantitative doxastic states presented in section IV

The reader can still ask What is the place of the Ramsey test for probability conditionals in the space of current theories of conditionals? The test is a garden-variety of "Ramsey test" recently used in the theory of epistemic conditionals[59] But it is a very special type of Ramsey test By the same token, one can ask what is the relationship between the test and the orthodox versions of probability logic Some of these questions have already been considered above, at least partially, but I now want to review the complete answers

---

[58] A careful analysis of the interplay of the notions of strict (positive) and negative validity in the context of the theory of epistemic conditionals is presented in Levi and myself, "Two Notions of Epistemic Validity," *Synthese*, CIX, 2 (November 1996) 217-62

All variants of Ramsey tests determine acceptability with respect to an epistemic state *E*, which can (but need not) be a proposition $a > b$ is accepted with respect to *E* if *B* is accepted in the hypothetical transformation of *E* with *A* Different versions of the test are determined by different choices of the underlying epistemology, which, in turn, determines the nature of *E* as well as the nature of the suppositional operation on *E* Therefore, different types of conditionals can be parametrically classified by varying the background epistemology used in the formulation of their acceptance conditions The test presented above is importantly constrained by an underlying epistemology of probabilistic kind, where the only primitive is a notion of generalized conditional probability Once this epistemology is adopted, the improved Adams hypothesis for acceptance can be formulated in purely qualitative terms, but this formulation induces a strong constraint on the nature of the epistemic state with respect to which acceptance is judged The intended targets are the *expectations* for the primitive two-place function *P*, rather than the full beliefs for that function The expectations for *P* are also the body of qualitative attitudes that need to be hypothetically transformed in order to evaluate probabilistic conditionals On the other hand, the underlying probabilistic epistemology imposes important constraints on this notion of supposition, which other nonprobabilistic operations on primitively defined expectations do not need to assume In section VII, I shall axiomatize the relevant notion of hypothetical change and compare it with other characterizations of supposition in the literature I shall also show that this notion of change is inconsistent with many of the well-known notions of supposition used in the literature The singularity of the Ramsey test for probability conditionals is in part determined by the singularity of the notion of supposition induced by the underlying probabilistic epistemology

I now turn to probability logic The cornerstone of the contemporary theory of probability conditionals is the idea (the *Adams hypothesis*) that probability of conditionals is conditional probability Perhaps the central lesson of Lewis's "Probabilities of Conditionals and Conditional Probabilities" (*op cit*) is to make clear that the Adams hypothesis is incompatible with the idea that conditionals express propositions So, there is a clear choice between the Adams hypothesis and truth conditionality Adams and his associates have chosen to give up truth conditionality in order to preserve the hypothesis The choice is motivated by the plausible idea that the assertability of conditionals goes by conditional probability This insight is a highly entrenched tenet of probabilistic semantics Truth conditionality, on the other hand, is a thesis which probabilists are ready to abandon on the basis of recalcitrant evidence [60]

---

[59] See Cross and Nute

Here we should face an important issue concerning my present proposal As Adams has made rather clear in recent writings, the important problem in probabilistic semantics is not *only* how to capture the inference patterns validated by probabilistic conditionals Many rival theories might be able to satisfy this requirement The crucial problem is to produce a theory capable of capturing these patterns *as well as* the insight that assertability of simple (and contingent) conditionals goes by conditional probability Adams made this point in *A Primer of Probability Logic* while commenting on Stalnaker's possible-worlds semantics [61]

> While the Stalnaker theory has the same consequences as the probabilistic theory so far as the validity of inferences is concerned, the probabilities assigned to conditionals by it are not conditional probabilities, and therefore it disagrees with the probability conditional theory at the level of probability (*op cit*, p 196)

Since Stalnaker assumes truth conditionality, the probability of his conditionals cannot go by conditional probability And this is so even when there is a coincidence at the level of validity for inferences involving simple conditionals

The notion of strict validity used here is also capable of capturing the inferential patterns sanctioned by the probabilistic theory But unlike rival views, there is no disagreement with the probability conditional theory at the level of probability. The present proposal follows the probabilistic theory in not assuming truth functionality Therefore, it is perfectly compatible with the idea that the degrees of assertability of conditionals go by conditional probability—to the extent that this idea can be coherently formulated

---

[60] The dominant view among contemporary probabilists is a moderate agnosticism about truth conditionality Perhaps there is a coherent theory showing that conditionals are truth bearers, but such theory is not needed in order to propose assertability conditions for conditionals understood as syntactical objects Other scholars from Ramsey to Allan Gibbard have been more emphatic in *denying* that conditionals express propositions

[61] Stalnaker's account is the one presented in "A Theory of Conditionals," in Nicholas Rescher, ed, *Studies in Logical Theory* (Oxford Blackwell, 1968), pp 98-112

It is important to notice that, although the acceptability of $a > b$ is determined by $P(B|A)$ assuming the value 1, my theory does not appeal to *decisive* functions with values ranging exclusively in $\{0, 1\}$ McGee[62] has recently used these functions in order to capture Stalnaker's logic via a probabilistic model Decisiveness is permitted but not obligatory in our models The crucial difference between models like McGee's and my present account is the appeal to a slightly modified version of van Fraassen's theory Corpora of expectations and full belief are determined for all functions in the model, including nondecisive ones

VI 2 Ramsey test for full belief The qualitative test for simple probability conditionals can be formulated for other attitudes rather than expectations

> *Ramsey test for two-place functions* a simple conditional $(a \Rightarrow b)$ is accepted with respect to $P(—|—)$ if and only if the largest core of $P[A]$ entails B

According to this test, the conditional $a \Rightarrow b$ is accepted with respect to a two-place function P if and only if as many as possible of the *full beliefs* for P as are compatible with A entail B So, the Ramsey test for two-place functions is closer in spirit to the various acceptance tests recently considered in the literature devoted to *epistemic conditionals* (where tests for acceptance of conditionals are usually formulated in terms of suppositional operations on bodies of full belief) In addition, the conditional patterns induced by the Ramsey test for two-place functions diverge from the ones validated by the Adams hypothesis for acceptance Here is an example to substantiate this claim The conditional $(True > \neg a) \rightarrow (a > False)$ is not strictly valid, while $(True \Rightarrow \neg a) \rightarrow (a \Rightarrow False)$ is Even when the acceptance of $(True > \neg a)$ determines that the unconditional probability of A is zero, only the acceptance of $(True \Rightarrow \neg a)$ is sufficient to determine the *serious impossibility* of A, in the sense that $P[A]$ is abnormal Acceptance of $(True > \neg a)$ is perfectly compatible with the normality of $P[A]$ This reinforces the claim made above to the extent that probability conditionals are, in an important sense, expectation conditionals[63] To be sure, one can study the logic of the conditionals induced by the Ramsey test for two-place functions, but these conditionals have a different logic from the one studied in probability logic

---

[62] Actually, McGee's result captures a non-nested fragment of Stalnaker's logic

VII HYPOTHETICAL REVISION

We are now in a position to see the main properties of qualitative supposition induced by two-place functions Consider a model $M = \langle S, C, \langle Ex, F\rangle, Sup\rangle$, the following postulates are satisfied by non-coreless functions P

Entailment $Ex(P) \subseteq F(P)$
Full belief expansion $F(P) \cap A = F(P[A])$
Success $Ex(P[A]) \subseteq A$
Preservation if $Ex(P) \cap A \neq \emptyset$, then $Ex(P) \cap A = Ex(P[A])$
Restricted consistency preservation if $A \cap F(P) \neq \emptyset$, then $Ex(P[A]) \neq \emptyset$
Entertainability if $F(P) \cap A = \emptyset$, then $P[A]$ is abnormal
Fixity if P is the abnormal function, then $Ex(P[A]) = F(P[A]) = \emptyset$, and $P[A] = P$
Cumulativity $Ex(P[A, B]) = Ex(P[A \cap B])$

Some of these postulates follow from basic definitions, others are corollaries of previous results For example, both the proof of cumulativity and the proof of preservation require substantial use of observation one, while restricted consistency preservation is a consequence of the lemma of descending chains The notion characterized by the previous postulates will be called from now on *hypothetical revision*[64]

It is worth noticing that full belief expansion, together with fixity, entails the converse of entertainability, which we can call *consistency* In other words, the nonentertainability, or serious impossibility, of a proposition A for P (that is, the fact that $F(P) \cap A = \emptyset$) is both a necessary and a sufficient condition for the abnormality of $P[A]$ While consistency is a widely accepted Bayesian principle, most qualitative theories of belief change allow for the formation of policies for changing one's view upon learning (supposing) that A when the agent is certain that A is false The probabilistic tradition based on the use of Popper functions is not equipped to deal with such full-belief-contravening changes

---

[63] It is also important to realize that no preferential (nonmonotonic) logic of the type studied by Lehmann and Magidor (*op cit*) validates the pattern $(true > \neg a) / (a > false)$ This also reinforces the idea that most nonmonotonic notions of consequence can be represented as *expectation conditionals* where the term 'expectation' is technically understood in probabilistic terms

[64] See my "Hypothetical Revision and Matter-of-fact Supposition," as well as Thomason and myself

So, one can arrive at the former characterization of probability-based belief change by defining $P[A](X|Y)$ as $P[A](X|Y \cap A)$, and by focusing on the properties of core systems. Alternatively, hypothetical revision can be adopted as a primitive, without specifying the meaning of $P[A](X|Y)$ via an explicit definition. It is easy to see that there is more than one notion of probability change compatible with hypothetical revision. On the other hand, there are quite robust properties entailed by the axioms just proposed. For example, the following property is entailed just by "full belief expansion":

Cumulativity for full belief $F(P[A, B]) = F(P[A \cap B])$

Cumulativity for full belief is an important qualitative property induced by the type of radical probabilism under consideration here. In fact, there is evidence that cumulativity is a central ingredient in the research program whose only primitives are either infinitesimal probability or two-place functions. Thomason and I (*op. cit.*) provide such evidence by showing that cumulativity for full belief is needed in order to extend to the iterated case the mapping between infinitesimal probability and extended conditional probability first developed by McGee.

All the theoretical pieces of the puzzle are now in place in order to substantiate one of the main claims made in the introduction above. One of the challenges presented there was to show that the so-called export-import law is a deeply entrenched commitment of the form of radical probabilism used in order to improve Adams's original hypothesis.

export-import $\phi > (\psi > \chi) \leftrightarrow (\phi \wedge \psi) > \chi$, where '$\leftrightarrow$' is the standard biconditional and $\phi, \psi \in L$, $\chi \in LC$.

The connections between cumulative models, exportation, and importation are articulated in the next section, together with other properties of probability conditionals induced by hypothetical revision.

VII.1. Exportation and importation. It is easy to see that the export-import law is strictly validated in the class of normal and countable models. This is guaranteed by the coincidence lemma established in section v (and proved in footnote 51), *and* by the definition $P[A](X|Y) = P(X|Y \cap A)$.[65]

---

[65] To see that notice that the acceptance of $a > (b > c)$ with respect to P depends on whether C is accepted with respect to $(P[A])[B]$. Since by definition $(P[A])[B](X|Y) = P(X|Y \cap A \cap B)$, we have that $a > (b > c)$ is accepted with respect to P as long as $P(C|A \cap B) = 1$.

The result can be generalized in an interesting manner In fact, two assumptions used above can be considerably weakened On the one hand, both exportation and importation hold in a larger class of uncountable models The following section will focus on those models On the other hand, the export-import law is valid in all cumulative models even when $P[A](X|Y)$ is not defined as $P(X|Y \cap A)$ This second point is particularly important given the centrality of cumulativity for the versions of radical probabilism under consideration It is clear, as we saw above, that cumulativity is enforced if $P[A](X|Y)$ is defined as $P(X|Y \cap A)$ But there might be alternative definitions of $P[\ ]$ whose associated notion of probability-based change are cumulative Both exportation and importation hold for all those definitions of $P[\ ]$ [66]

Although I am not focusing here on the conditionals induced by the Ramsey test for full belief, I can say in passing here that the export-import laws also hold for those conditionals In fact, cumulativity for full belief is a rather robust feature of the semantic program based on the use of generalized conditional probability Of course, $a > a$ (success) is immediately validated, and a straightforward consequence of having the export-import law and success is (global success) $a > (b > a)$ It should also be mentioned that neither export-import nor global success is compatible with most of the standard theories of theory change Intuitively, global success says that every input hypothetically accepted at a certain point is rigidly maintained in future changes This is against the one-shot property of consistency preservation assumed by almost all theories of theory change The idea of consistency preservation is that a consistent corpus changed by a consistent input should yield a consistent output It is easy to see that the method induced by two-place functions is not consistency preserving (just consider $a > (\neg a > a)$) It is true, nevertheless, that theories of conditionals developed with an eye on the role of conditionals in hypothetical reasoning and decision making tend to assume variants of export-import without much hesitation (sometimes for mathematical convenience) For example, Richard Bradley[67] offered one of those theories in a recent article The present model validates all of his axioms (*ibid*, p 191) with the exception of the schema $\neg(a > b) \leftrightarrow (a > \neg b)$ This axiom is rather controversial, and Bradley accepts it only reluctantly My model validates a weaker property in the presence of normality (fullness) $\neg(a > \gamma) \leftrightarrow (a > \neg\gamma)$, where $\gamma \in LC$ and $a \in L$

---

[66] This can be proved directly In fact, cumulativity guarantees that assuming the antecedent of each law yields the correspondent consequent Notice that here we are assuming cumulativity as an axiom rather than relying on a particular definition of $P[\ ]$ —and then using observation one

[67] "A Representation Theorem for a Decision Theory with Conditionals," *Synthese*, CXVI, 2 (1998) 187-229

VIII INFINITE PROBABILITY SPACES

So far I have mainly focused on countable spaces We have seen that certain conditional laws continue to be valid even if the countable restriction is lifted Nevertheless, there are many reasons for being very cautious in incorporating uncountable spaces Some of these reasons are technical and go deep into the foundations of probability theory The gist of the problem is that according to the received view in conditional probability, if w is a point in U, $P(—|w)$ is not probability given an *event*, but probability given a sigma field Some authors have formulated variants of the theory of two-place functions offered here, where relativizations are used For example, Renyi's characterization of conditional probability does appeal to a relativization, not to sigma fields but to *bunches* Formally, B is a *bunch* of a space S if it is closed under finite unions, the empty set does not belong to B, and there is a sequence of events in B whose union equals the universe U of S Intuitively, the bunch should group a set of *observable* events, and only those events should be admissible conditioning events Unfortunately, Renyi does not offer a clear operational test for identifying observable events, although some researchers have recently proposed such tests in particular areas of application of the theory We can easily adopt Renyi's axiomatization here, but ultimately this does not seem sufficient to produce a smooth connection with contemporary measure theory This is so, at least, when countable additivity is adopted and the space is uncountable There is, nevertheless, no need to go deeper into this issue here In fact, I shall suggest below that (1) there seem to be good reasons for not focusing on uncountable models in the first place, and (2) the axiom of countable additivity is suspect for the application we are considering

Although the gist of Adams's hypothesis is conceptually appealing, there is no guarantee that it will continue to be conceptually compelling when applied to uncountable probability spaces The notion of probability delivers a nice account of the notion of measure and size comparison, and it does so both in the infinite and finite case But it is not clear that an epistemological interpretation attached to measures (as encoding degrees of belief or degrees of conditional belief) needs to be robustly maintained for spaces of different sizes In fact, I would like to offer reasons against that view by defining an infinite conditional probability space, and proposing a concrete example showing that the probability function so defined delivers unintuitive acceptance conditions.

Renyi provided several examples of conditional probability spaces obeying both the reduction axiom and the multiplication axiom introduced in section IV. For example, let R be the real line, and let BR be the σ-algebra of Borel subsets of R. Let $\lambda(A)$ denote the Lebesgue measure of a set $A \in BR$. For all A and B in BR such that $\lambda(B)$ is strictly greater than 0 and strictly less than positive infinite, we can define $p(A|B)$ as $\lambda(A \cap B)/\lambda(B)$. We then get a *full* probability space. This space does obey the reduction axiom and the multiplication axiom. Consider now the following situation in which an agent evaluates the weight of a piece of stone in the real interval [0,1]. We can assume that the function p encodes the agent's epistemic state. Notice that, according to the above definition, the agent is committed to accept 'If a number measures the weight in the [0,1] interval, then the number in question is irrational'. There are plenty of examples of this kind—for example, 'If the pointer is set up in such a way that it might land randomly at any point in the real interval [0,1], then it will land at an irrational number'.

IX. COUNTABLE ADDITIVITY AND EPISTEMIC PARADOX

Up to this point, I have taken for granted that the underlying one-place probability functions obey Kolmogoroff's axioms. Is there any reason for thinking that these classical axioms (stated in section IV) might be problematic? Here I would like to argue that countable additivity might be questionable.

One of the main goals of van Fraassen's framework is to deal with the so-called *lottery paradox*: say that agent *a* assumes that the weight of a stone is representable by a real number in some interval, say, between .5 and 1 pounds. *a* is certain of the hypothesis that stipulates that the weight lies somewhere in the interval. Yet *a* might coherently assign 0 probability to the uncountable many rival hypothesis of the form, 'The weight is exactly x pounds, where x is a real value between .5 and 1'. So, *a* fully believes that each of these rival hypothesis will not obtain, but he is also certain that one of them will obtain.

van Fraassen's construction circumvents this type of situation by denying that probability 1 is sufficient for full belief. Consider now the following variant of the paradox of the lottery: take any effective enumeration $r_1, r_2, \ldots, r_n, \ldots$ of the rationals in [.5, 1]. Consider now the following family T of hypothesis: (*T1*) the weight is exactly r_1, (*T2*) the weight is exactly r_2, ..., (*T3*) the weight is exactly r_n, ...

There are countably many of these hypotheses. It seems excessive to dismiss as irrational the idea that the hypotheses in T are equiprobable Now, if the hypotheses in T are equiprobable, they cannot carry positive probability (because finite additivity guarantees that some finite union of hypotheses in T will carry more than unit probability) But one can assign 0 probability to each T_i Then all the probability mass will concentrate in co-finite families of hypothesis in T But now the assumption that probability 1 entails full belief leads to a transfinite lottery paradox For one should fully believe that each of the countably many rival hypothesis T_n will not obtain, and one should also be practically certain that the weight lies somewhere in the rational interval [5, 1]

Symmetry seems to require a uniform treatment of both countable and uncountable transfinite lottery paradoxes van Fraassen's theory does account for both paradoxes, but the account is asymmetric The first paradox is resolved because in van Fraassen's theory, probability 1 does not automatically entail full belief The treatment of the second paradox is slightly more ad hoc Since van Fraassen's unconditional probability functions obey countable additivity, the scenario depicted in the former example is simply ruled out as incoherent Of course, this is also a problem for the improved Adams hypothesis In fact, consider (*EQ*) if the pointer is set up in such a way that it might land randomly at any point in the rational interval [0,1], then the possible outcomes are equiprobable It seems too extreme to rule out a priori the acceptability of (*EQ*) as a form of epistemic incoherence Nevertheless, this is exactly what the improved Adams hypothesis recommends, when the underlying unconditional probability is countably additive To see that, notice that the proposition $E = \{x \in U\ P(\{x\}|U) = 0$ and for every point $y \in U$ such that $y \neq x$, $P(\{y\}|U) = 0\}$ is, according to the axioms, empty

IX 1 Finitely additive models A radical solution to the problem presented above is just to abandon countable additivity In that case, the adoption of a uniform unconditional prior on a countable space is perfectly possible And, of course, *EQ* ceases to be analytically false with respect to finitely additive models

Many crucial features of van Fraassen's picture remain untouched when we adopt finite additivity For example, each non-coreless function continues to induce a nested system of cores Nevertheless, there is no guarantee of the existence of an innermost core Therefore, the definition of expectations cannot be done in terms of the intersection of all cores But this is not an obstacle against defining expectations One can appeal to the type of definition used by van Fraassen himself One can say that the proposition A is expected (for

P) if and only if A is either a priori or entailed by some core for P[68] So, for non-coreless P, we can define $Exp(P)$ as the set of all propositions expected for P

$$Exp(P) = \{A \quad A \text{ is entailed by some core of } P\} = Exp(P)$$

This move still allows us to characterize the acceptance of simple conditionals qualitatively in terms of what is entailed by expectations Basically, $a > b$ is accepted with respect to P if and only if B is in $Exp(P)$ And, of course, the definition of full belief remains unchanged Nevertheless, an interesting symmetry is immediately broken In the countable case, $P[A](B) = 1$ guarantees that some core for $P[A]$ entails B Once countable additivity is abandoned, it is possible that there is no core for $P[A]$ entailing B although $P[A](B) = 1$ As an example, assume that $P(B|A) = 1$ and set $B \subseteq A$ The following situation can then arise One can have infinitely many cores C (for P) cutting B (but not included in B) in such a manner that the intersection of each C with B carries probability 1 So, the following test does not capture the qualitative counterpart of the Ramsey test for probability conditionals when countable additivity is abandoned

> *Expectation test for probability conditionals* $a > \beta \in Sup(P)$ if and only if (1) $b = \beta$ is in L, and $B \in Exp(P[A])$, or (2) β is in LC, and $\beta \in Sup(P[A](X|Y)) = Sup(P(X|Y \cap A))$

Of course, the expectation test is indistinguishable from the Ramsey test when countable additivity is enforced This means that the fact that $P(A|B) = 1$ is in this case perfectly represented by the fact that B *is expected (for P) given A* But, once we consider finitely additive models, $P(A|B) = 1$ is only a necessary but not a sufficient condition for B being expected given A (for P) It is worth pointing out that the expectation test might have its own separate interest In fact, the same rationale that motivates the use of the Ramsey test for probability conditionals motivates the basic idea behind the test[69] As before, when $F(P) \cap A = \varnothing$, we set $Exp(P[A])$ to $\varnothing$ By the same token, the expectation of the abnormal two-place function is set to $\varnothing$

The export-import laws are a rather robust part of the probabilistic program In fact, *they continue to be validated under either test* A final issue remains open Is there a new qualitative condition representing P($A|B$) = 1 for finitely additive models? The answer is "yes " In fact, for non-coreless functions P we can define

---

[68] van Fraassen did appeal to this formal maneuver in order to define full belief in "Fine-grained Opinion, Probability, and the Logic of Full Belief"

[69] The exact formal details of the conditional logics induced by the expectation test (and by the generalized test) remain an open problem in the field

K is a *kernel* of a core C for function P if and only if $K \subseteq C$ and $P(K|C) = 1$

So, for non-coreless P, we can define the *probabilistic kernel* for P, as the set of all propositions entailed by some kernel of some core for P $PK(P) = \{A$ A is entailed by some kernel of some core of $P\}$

> Observation (2) if the universe U of the space S is normal for a function P defined on S, and A is epistemically possible for P, $P(B|A) = 1$ if and only if $B \in PK(P[A])$ [70]

Although we can easily define acceptance conditions in terms of what is entailed by probabilistic kernels, the idea seems to have less intuitive interest than the one based on the expectation test A curious situation thus arises where an acceptance test, which formally departs from Adams's test (the expectation test), might, however, preserve in a better manner the intuitions that motivate probabilistic acceptance tests on qualitative grounds—that is, in spaces of at most countable size The complete formal details of the conditional systems validated by each test are still unknown My main point in this section was to stress the fact that export-import is a salient part of the program in probabilistic semantics, even when we consider large classes of models and important weakenings of the axiomatic base used to characterize both conditional and unconditional probability

HORACIO ARLÓ-COSTA

Carnegie Mellon University

---

[70] Observation (2) can be easily established as follows assume first $B \in PK(P[A])$ Therefore, there is a core C such that it has a kernel K entailing B So, $P[A](K) = 1 = P(K|A)$ Finite additivity and the inferred normality A gives us then that $P(B|A) = 1$ Assume now that $P(B|A) = 1$ We have to consider two main cases First assume that B does not cut any core C for $P[A]$ This leads to a contradiction because $P[A](C|U) = 1$ for an arbitrary core for P Therefore, $P[A](C + B|U) = 2$ by finite additivity and the assumed normality of U Assume now that there is a core C for $P[A]$ such that $C \cap B \neq \varnothing$ Then, $C \cap B$ carries probability 1 according to $P[A]$ For assume that there is Z in C such that it does not cut $C \cap B$ and $P[A](Z|U) > 0$ Then, by finite additivity, $P[A](B + Z|U) = P[A](B|U) + P[A](Z|U) > 1$, taking into account that we assumed $P[A](B|U) = 1$ Since $P[A](C|U) = 1$, axiom II guarantees that $C \cap B$ is a kernel for C for the function $P[A]$, and this kernel entails B

COMMENTS AND CRITICISM

WHEN "EVIDENCE TRANSCENDENCE" IS NOT MALIGN
A REPLY TO CRISPIN WRIGHT

As always, I learned from Crispin Wright's[1] article, and I am sure that the points on which we still disagree will provide fruitful material for future discussions. In many ways, Wright is the reader I had in mind when I wrote the papers and books he discusses. (In what follows, I have found it easiest to discuss Wright's points in roughly the reverse of the order they occur in his article.)

I begin with a quote from the final paragraph:

> It is essentially the interface picture that we will need to make sense of the idea that 'There are no intelligent extraterrestrials' might be true beyond all evidence available in principle, beyond the verdict of even the very best possible empirical theory (364).

To explain why I do not agree with this, let me go back a few paragraphs to some remarks in which Wright lays out in a clear and very general way the position for which he is arguing. He makes the point that there are benign forms of evidence transcendence:

> This benign form of evidence transcendence is conceived as going with—in the broadest sense—*contingencies of epistemic opportunity*: in all cases where we have a conception of this kind of how the truth value of a particular statement could be unverifiable, a developed specific account of that conception will consist in detailing limitations of opportunity, or spatiotemporal situation, or perceptual or intellectual capacity, which stop us getting at the relevant facts but to which we, or others, might easily not have been subject—or at least, to which we can readily conceive that an intelligible form of investigating intelligence need not be subject (360).

In the paragraphs which follow, Wright contrasts this benign form with the "malign" kind of evidence transcendence posited by interface conceptions:

> On this type of view, evidence transcendence is a product not of contingencies of opportunity, or of contingent limitations of our powers, but of *metaphysical* shortcomings and divides: consciousness—unless it is

---

[1] "Truth as Sort of Epistemic: Putnam's Peregrinations," this JOURNAL, XCVII, 6 (June 2000): 335-64.

0022-362X/01/9811/594–600

> the consciousness of God, of which we have no satisfactory conception—is necessarily and essentially insular
>
> The essence of metaphysical realism, we might say, is thus *interface realism* And the evidence-transcendent conception of truth which metaphysical realism brings in train is of a malignant kind, the kind that goes with an interface conception of mind's interaction with the world About these ideas, there is nothing "natural" or "commonsensical " Rather, it is supposed to take philosophical sophistication to appreciate their point (361)

I agree with Wright that the sorts of evidence transcendence that we can explain as resulting from "contingencies of epistemic opportunity" in the broadest sense should be admitted to exist by any philosophy that aims to do justice to the natural realism of the common person (as Michael Dummett's antirealism[2] about the past evidently did not) But unlike Wright, I believe that 'There are no intelligent extraterrestrials' might conceivably be true, that this too should be admitted by any philosophy that aims to do justice to common-sense realism, and that the evidence transcendence of the truth of this proposition should be classed as "benign" and not as "malignant "

To see why, let us compare it with a "Fitch-example" (Wright calls it $\mathscr{F}$) whose truth is also evidence transcendent, but in a way Wright considers "benign"

Q and no one will ever rationally believe that Q

(Here let Q be any empirical proposition that it might happen that no ever rationally believes, say, *There is a gold mountain one mile high* [3])

According to Wright, we can give a truth condition for $\mathscr{F}$, in the fashion of a Carnapian "reduction sentence" (Wright calls it a *provisional biconditional*) The condition is

(1) If $\mathscr{F}$ was to be appraised under sufficiently good circumstances, $\mathscr{F}$ would be true just in case it was believed to be true (357)

But there is a catch! As Wright notes, there are only two sorts of circumstances in which it is possible to "appraise" the truth value of $\mathscr{F}$ these are (1) circumstances in which it is possible to know that Q is false, and (2) circumstances in which it is possible to know that someone rationally believes that Q is true, and in both sorts of circumstances $\mathscr{F}$ is false (because either its first or its second conjunct is false) In sum, what (1) captures is the fact that $\mathscr{F}$ is falsifiable *But it is not verifiable* If we say that $\mathscr{F}$ might be true, we must presuppose our hearer's possession of an idea of what constitutes $\mathscr{F}$'s being true which (1) does not capture This would become crystal clear if the "sufficiently good circumstances" in (1) were spelled out "constructively," in Wright's sense, for then it would be obvious that the "circumstances" in question are all circumstances in which $\mathscr{F}$ is false A "reduction sentence," as Rudolf Carnap and Carl G Hempel[4] both noted, leaves the extension of the concept it "reduces" undetermined when the test condition mentioned in the antecedent is not fulfilled Wright's provisional biconditional is just a reduction sentence for the predicate 'true', one which partially specifies its extension as applied to $\mathscr{F}$, and it leaves the extension of that predicate undetermined in every case in which the "sufficiently good circumstances" do not obtain—including all the circumstances in which $\mathscr{F}$ is true

---

[2] My internal realism was never committed to *this* position of Dummett's, Dummett, "The Reality of the Past," in *Truth and Other Enigmas* (Cambridge Harvard, 1969)

[3] I used this very example, not knowing of F B Fitch's earlier argument, in "Logical Positivism and the Philosophy of Mind," in my *Mind, Language and Reality Philosophical Papers, Volume 2* (New York Cambridge, 1975), chapter 22

Thus one of the following two remarks by Wright must be wrong

> (a) " a moderate-internalist conception of truth [in Wright's sense] is friendly to classes of potentially recognition-transcendent truths, typified by Fitch examples" (358)
>
> (b) " in all cases where we have a conception of this kind of how the truth value of a particular statement could be unverifiable, a developed specific account of that conception will consist in detailing limitations of opportunity, or spatiotemporal situation, or perceptual or intellectual capacity, which stop us getting at the relevant facts but to which we, or others, might easily not have been subject—or at least, to which we can readily conceive that an intelligible form of investigating intelligence need not be subject" (360)

To spell this out if Wright's "moderate internalist conception of truth" does indeed allow that some Fitch propositions can be true (I take this to be what is meant by the reference to the "friendliness" of the conception to this class of potentially recognition-transcendent truths), it had better not assert (b)! For it is not the case that our conception of how a Fitch proposition might be true (as opposed to our conception of how it might be false) enables us to say how "we, or others" might have been able to recognize its truth

---

[4] Hempel, "The Empiricist Criterion of Meaning," in *Revue Internationale de Philosophie,* IV (1950), collected in A J Ayer, ed , *Logical Positivism* (New York Free Press, 1959), pp 108–29, see, for example, p 120

To make matters even stranger, the situation is exactly the same if instead of $\mathcal{F}$, which Wright sees as exemplifying "benign" recognition transcendence of truth, we take $\mathcal{N}$: "There are no intelligent extraterrestrials," whose (possible) recognition-transcendent truth Wright sees as "malign." For that proposition is falsifiable-though-not-verifiable just as $\mathcal{F}$ is. Its truth value, like $\mathcal{F}$'s, can be "appraised under sufficiently good circumstances"—circumstances in which it is false! For if $\mathcal{N}$ is false, then there are intelligent extraterrestrials, and we may easily imagine there being observers who know this. We can even formulate a 'provisional biconditional' for $\mathcal{N}$—simply replace '$\mathcal{F}$' by '$\mathcal{N}$' and 'believed' by 'verified'!

This brings us to the question: Given that he does not think that "Fitch examples" involve any sort of interface conception, why does Wright say that "It is essentially the interface picture that we will need to make sense of the idea that 'There are no intelligent extraterrestrials' might be true beyond all evidence available in principle, beyond the verdict of even the very best possible empirical theory" (365)? Earlier in the article, he writes:

> If moderate internalism can accommodate the example [our $\mathcal{N}$] as a truth-apt claim, then the view has to be that under such informationally fortunate circumstances [knowledge of the conditions for the evolution of intelligent life, and "enough understanding of the distribution and diversity of matter in the cosmos to allow the assignment of a well-grounded probability to the supposition that such conditions are replicated elsewhere besides on this Earth"], thinkers would take the correct view of the question (353).

It is the idea that thinkers would not take the correct view of the question under such circumstances that Wright analogizes to belief in a metaphysical "interface." But this is a confusion.

To see the confusion, let us assume, as is reasonable on present knowledge, that the "very best possible empirical theory" that Wright speaks of tells us that the cosmos, or at least the part of space-time that might contain intelligent life, is finite, and that there is no one place within that finite region from which one can survey (or even receive causal signals from) the entire region. Let us suppose, with Wright, that the theory, call it *P*, provides "enough understanding of the distribution and diversity of matter in the cosmos to allow the assignment of a well-grounded probability to the supposition that such conditions are replicated elsewhere besides on this Earth," and let us even suppose that the probability in question is quite high, say, .95. Now, it is a conceptual truth (what Wright calls an "a priori" truth[5]) that what is highly probable does not always happen Thus it is internal to physical theory itself (to P)—and not the product of some metaphysical conception of "the kind that goes with an interface conception of mind's interaction with the world"—that $\mathcal{N}$ may be true If "moderate internal realism" requires us to believe that all those propositions which our best theory regards as highly probable (but no more than that) are in fact true, then "moderate internal realism" must certainly be rejected!

What then *is* the source of the recognition transcendence of true Fitch propositions and true negative existentials similar to $\mathcal{N}$ (see *ibid*)? The answer I offer both in the Dewey Lectures[6] and in an article I believe Wright may have missed,[7] is that while it is indeed the case that our ability to verify them perceptually under favorable circumstances is part of what is comprised in our understanding of such atomic descriptive propositions as 'Here is a chair' (and thus I do not need to be persuaded by Wright that we should not "recoil" so far in the direction of metaphysical realism as to think of truth as recognition transcendent in the most ordinary cases), the use of logical connectives and quantifiers does frequently lead to recognition transcendence for reasons that are clear from the logical structure itself (Note that even the Fitch examples involve negative existentials!) Nor is the principle of bivalence what is involved here Just as the logical structure of a Fitch example makes it clear that what is supposed in such an example—and I see no reason to regard it as inconceivable, or "metaphysical," or "confused," and so on—is something that, if it happens, we cannot verify, so the logical structure of $\mathcal{N}$ makes it clear that what is supposed in this example—say, that something improbable (if it *is* improbable) simply happens—is something that, if it happens, we cannot verify[8] Both sorts of example

---

[5] My qualms about "a priori" stem from the long-standing associating of the notion with the idea of unrevisability, not with any rejection of the idea of conceptual analysis Cf my "Pragmatism," *Proceedings of the Aristotelian Society*, XCV, 3 (1995) 291–306, especially pp 299-302

[6] These were originally published as "Sense, Nonsense, and the Senses An Inquiry into the Powers of the Human Mind," this JOURNAL, XCI, 9 (September 1994) 445-517, they are collected in my *The Threefold Cord Mind, Body and World* (New York Columbia, 2000)

[7] See "Pragmatism "

[8] More precisely, the conjunction P & N & S, where S is the statement that an observer cannot verify whether there is intelligent life in a region of space time that that observer is unable to receive causal signals from, P is the "possible empirical theory" that tells us that (1) causal signals do not travel faster than light, (2) it is physically possible (and highly probable) that there are intelligent extraterrestrials, but (3) it is also physically possible that there are not, and (4) there are large regions of space time that any particular physically possible observer is unable to

(and of course there are others) spell disaster for any elucidation of truth in terms of rational acceptability, even by means of reduction sentences (provisional biconditionals)

Much of Wright's paper is concerned with interpretation of and improvement on the theory of truth presented in my *Reason, Truth and History*[9] (a theory I no longer accept for the reasons given in my Dewey Lectures) I shall venture only three brief remarks on what I believe I meant in the insufficiently clear passage with which Wright wrestles

(1) What I had in mind was the following for a statement *P* to be true (on the "internalist account"), there must be a set of conditions *C* which are "good" (say, if we indulge in the fiction that there are numerical degrees of goodness, to some degree *g*) for appraising *P*, such that *P* is assertible under conditions *C*, and such that it is the case that if conditions C' were still better (say, the degree of goodness for appraising *P* were g' instead of *g*) then *P* would still be assertible under C' The principle governing all use of idealizations in solving practical problems (hence the example of "frictionless planes") is that we can treat conditions that are close to (but not the same as) some ideal conditions *I* as if they were the ideal conditions (as if the incline were frictionless) if improving the approximation would not affect the result I believe, in fact, that my internalist notion of truth coincided with the notion that Wright calls "superassertibility" in his *Truth and Objectivity*[10] (although that is not how *he* interprets me!), if we prescind from possible differences over the notion of "assertibility "

(2) I did not and do not think, however, that one can survey in advance what one might reasonably come to understand by "sufficiently good epistemic conditions," even for a particular topic Certainly, there is no fixed set of "standards" associated with language which could enable one to do this (This is the possible difference over the notion of "assertibility" to which I just referred)

(3) Wright is quite correct in supposing that I did not consider the problems posed by the conditional fallacy in any of my formulations I agree that they are serious—indeed, I do not think that provisional biconditionals can meet them either (precisely because, like all reduction sentences, they leave the extension of the predicate that they

---

receive causal signals from, is a statement that it is logically impossible to verify if true, and yet it is a statement that *P* tells us could be true (and even assigns a probability to)

[9] New York Cambridge, 1981

[10] Cambridge Harvard, 1992

"elucidate"—Carnap's term was 'reduce'—undetermined with respect to cases in which the test condition is not fulfilled, and this is precisely what we do not want in the case of recognition-transcendent truths) I at least want to say that, for example, $\mathcal{N}$ can be true in such cases, not that it is indeterminate Even in the case of truths which are not recognition transcendent but simply such that (contingently) the test condition is unfulfilled, it is not enough to specify via a provisional biconditional what the truth value would have been if the recognition condition *had* been fulfilled, for this leads right back to the problems of the conditional fallacy My conclusion is that the problem Wright poses for my former view is also an (unsolved) problem for Wright's own "moderate internalism "

HILARY PUTNAM

Harvard University

BOOK REVIEWS

A Parting of the Ways Carnap, Cassirer, and Heidegger MICHAEL FRIEDMAN Chicago Open Court, 2000 xv + 175 p Cloth $49 95, paper $24 95 [*]

Michael Friedman has written an eye-opening and ambitious monograph on three exceptional figures in twentieth-century philosophy Eye-opening, because he offers us a significantly new perspective on the split between "analytic" and "continental" philosophy Ambitious, because he combines wide-ranging historical scholarship with a bold attempt to spell out an entire philosophical agenda

Friedman's main thesis is straightforward He holds that the nature and sources of the analytic/continental divide can be greatly illuminated by seeing that the two traditions have "evolved in sharply diverging directions from a common neo-Kantian heritage" (xi) Friedman's reminder of the crucial role that neo-Kantianism played before the rise of both analytic and continental philosophy is valuable The neo-Kantians prepared, in fact, the ground for much of what went on in the philosophy of the twentieth century But their contributions to philosophical logic, to the philosophy of mathematics and the philosophy of science, their interest in the cultural sciences and in ethics, and their creative concern with the history of philosophy have for the most part been forgotten Friedman is rightly convinced that the neo-Kantian tradition deserves our fullest attention, if we want to understand what has gone on since then

This he seeks to illustrate by depicting the early work of Rudolf Carnap and Martin Heidegger as reacting in opposite directions to the neo-Kantian thematic He writes

> Just as Carnap begins his philosophical career by attempting to realize the philosophical ambitions of the Marburg School of neo-Kantianism using the new mathematical logic created by Frege, Heidegger begins his career by attempting to resolve the outstanding problems of the Southwest School using the new phenomenological method due to Husserl (149)

This forthright assertion may, however, need qualification There is no question that all European thinkers in this period are reacting in

---

[*] I am grateful to Barry Stroud, Christopher Pincock, and Michael Friedman for comments on an earlier version

0022-362X/01/9811/601-11

some fashion or other to the Kantian heritage This is obvious not only for Carnap and Heidegger but just as much for Gottlob Frege, G E Moore, and Bertrand Russell, as well as for Edmund Husserl and Ernst Mach, and indirectly (via Arthur Schopenhauer) even for the early Ludwig Wittgenstein [1] But the ways and degrees in which they respond to the Kantian system and its neo-Kantian reconstructions differ remarkably The question is therefore to what extent neo-Kantianism is of specific significance for the development of Carnap's and Heidegger's thinking

Friedman draws in his discussion of Carnap on the fact that he began his philosophical studies under Bruno Bauch, a leader of the Southwest wing of neo-Kantianism, "who was also influenced by the more scientifically oriented Marburg School" (63) Carnap's examination of intuitive space in his dissertation is, moreover, according to his own testimony directly "under the influence of Kant and the neo-Kantians, especially Natorp and Cassirer" (cited 65) Friedman admits that after 1924 Carnap begins to distance himself from neo-Kantianism, but even in *The Logical Structure of the World* (1928),[2] "Carnap by no means intends simply to supplant neo-Kantianism by 'positivism'" (71) He asserts, rather, in §75 of his book that "the merit of having discovered the necessary basis of the constitutional system belongs to two entirely different, and often mutually hostile, philosophical systems" (cited 71) While positivism has stressed the basis of cognition in experience, neo-Kantians like Heinrich Rickert, Ernst Cassirer, and Bauch have emphasized the need for positing additional systems of ordering This much in Friedman's account is indubitable The question for the interpreter, however, is how deep the neo-Kantian influences go and how essential they are for the overall project of the *Aufbau* Friedman's account seems to conflict here with Carnap's own later claim, in his "Intellectual Autobiography,"[3] that "the strongest effect on my philosophical thinking were Frege and Russell" and that Frege influenced him most decisively in logic and semantics while "in my philosophical thinking in general I learned most from Bertrand Russell" (*ibid*, pp 12, 13) Carnap's "Intellectual Autobiography" also declares without qualification that the *Aufbau* was "inspired by Russell's description of the aim and method of future philosophy" (*ibid*, p 16) One might, of course, dismiss such characterizations as errors of memory, but the correspondence between Russell and Carnap during the 1920s fully confirms this later account Friedman's underestimation of the influence of Russell on the *Aufbau* and his resultant overestimation of that of the neo-Kantians appears to be due to a simplistic picture of Russell's position which he has taken over from W V Quine and Nelson Goodman According to this picture, Russell is just another empiricist "in the tradition of Locke, Berkeley, Hume, Mach," whereas closer study of his texts can show us that there exists, in fact, a significant overlap between Russell's and Carnap's projects in the period in question [4]

---

[1] Neo-Kantian influences on Frege have been explored in Gottfried Gabriel, "Frege als Neukantianer," *Kantstudien*, LXXVII (1986) 84–101, and my "Frege on Meaning," in Hans-Johann Glock, ed , *The Rise of Analytic Philosophy* (Malden, MA Blackwell, 1997), pp 17-34

[2] Rolf A George, trans (Berkeley California UP, 1967)

[3] In Paul Arthur Schilpp, ed , *The Philosophy of Rudolf Carnap* (La Salle, IL Open Court, 1963), pp 3-84

As far as Heidegger is concerned, Friedman proceeds from the premise that he "received his philosophical education within the neo-Kantian tradition of the Southwest School as it was articulated by Rickert in Freiburg" (39) In his doctoral dissertation and in his *Habilitationsschrift,* Heidegger showed himself therefore not surprisingly "to be a faithful follower of Rickert indeed" (40) After Rickert had left Freiburg in 1916, Heidegger became admittedly attracted to the more developed transcendental psychology of Husserl, Rickert's successor But "the problematic he inherited from Rickert and Lask helps to explain why Heidegger could not remain satisfied with Husserl's conception of phenomenology" (46) The result was the "subjective logic" of *Being and Time,*[5] "the so-called existential analytic of Dasein" (46) with which Heidegger finally thought he had overcome the problematic of neo-Kantian epistemology This picture of Heidegger's philosophical development differs sharply, however, from Heidegger's own in the *curriculum vitae* attached to his *Habilitationsschrift*—composed at the time when his professional future was entirely dependent on Rickert's good will He speaks there of Plato, Thomas Aquinas, Bonaventure, and textbooks of scholastic philosophy as the sources of his earliest philosophical thought There followed the reading of Husserl's *Logical Investigations,*[6] "which became decisive for the course of my scholarly development " Subsequently, in the winter of 1911, Heidegger began to attend Rickert's seminars at Freiburg, which introduced him to an understanding of the nature of logic and of modern philosophy since Kant Nevertheless, "my basic philosophical convictions remained those of Aristotelian-scholastic philosophy" The doctoral dissertation sought therefore to draw equally "on modern logic and basic Aristotelian-scholastic judgments" The reading of J G Fichte, G W F Hegel, and Wilhelm Dilthey and further study of Rickert led him from there to a new engagement with medieval philosophy Thus originated the *Habilitationsschrift* on theories of categories and meaning in John Duns Scotus whose wider aim was "a comprehensive account of medieval logic and psychology in the light of modern phenomenology" The impression of an intellectual distance to Rickert that Heidegger gives in these words is confirmed by Rickert's reaction to the *Habilitationsschrift,* though he was its official evaluator, he left the reading and assessment of the work, in fact, to a substitute, the Catholic theologian Engelbert Krebs [7] When Heidegger writes *Being and Time,* neo-Kantian themes seem, in any case, to have become entirely residual to him, what dominates are, instead, phenomenological, existential, and historicist strains of thought

---

[4] Friedman discusses the question of the significance of Russell for Carnap in *Reconsidering Logical Positivism* (New York Cambridge, 1999), which contains a more detailed exposition of his neo-Kantian reading of Carnap Drawing on the as yet unpublished Russell-Carnap correspondence, Christopher Pincock provides a decisive rebuttal of Friedman's interpretation—"Russell's Influence on Carnap's *Aufbau,*" *Synthese* (forthcoming)

[5] John Macquarrie and Edward Robinson, trans (New York Harper, 1962)

[6] J N Findlay, trans (New York Humanities, 1970)

These observations are naturally insufficient as a response to Friedman's subtle and informed neo-Kantian reading of Carnap and Heidegger, but they ought to suffice for showing that his interpretation is controversial Friedman recognizes, of course, that his account cannot be the whole story He freely acknowledges that his emphasis on the bearing of neo-Kantianism on Carnap's thought ignores "influences coming from Wittgenstein, Russell, the empiricist tradition, and even Leibniz," and that in Heidegger's case his account is "at the expense of the genuinely ontological preoccupations" that Heidegger has derived from the ancient Greeks (xii) But are such acknowledgments enough?

Even more problematic is his broader claim that the analytic/continental divide is best understood in terms of Carnap's and Heidegger's divergent responses to the neo-Kantian heritage His story ignores Moore's and Russell's revolt against the idealist tradition at the turn of the century This was recognized at the time as a "split" by all the parties involved, as we can see from R G Collingwood's[8] angry denunciations of the new generation of anti-idealist English philosophers Friedman's account also ignores the late nineteenth-century growth of a militantly antiphilosophical positivism in continental Europe which in Vienna crystallized first in the work of Mach and then in the first Vienna Circle of 1908 Here, too, there was from the start a clear consciousness of a "divide" between the fronts quite independent of what went on between Carnap and Heidegger[9] These episodes suggest that the parting of analytic and continental philosophy was certainly not due to a single event, a single dispute, a single issue but the outcome of a whole and diverse series of differentiations That picture corresponds to the realization that analytic philosophy has no single point of origin but is from the beginning a cooperative, collective, supranational enterprise[10]

---

[7] The *curriculum vitae* is reprinted in full in Hugo Ott, *Martin Heidegger A Political Life,* Allan Blunden, trans (New York Basic, 1993), pp 84-86

[8] *An Autobiography* (New York Oxford, 1939) His remarks are quite unaffected by what has been going on in continental Europe Russell and his followers, he writes, have jettisoned whatever can be recognized as a positive doctrine in philosophy and have simply revived "the old positivist attack on metaphysics" But their work, "with all its ingenuity and pertinacity, is only building card-houses out of a pack of lies"

---

(52) These are words as strong as any exchanged between Carnap and Heidegger They are understandable only in the light of Collingwood's feeling that, as a survivor of the idealist tradition, he is cut off and isolated from every other school of philosophy in England

[9] Friedrich Stadler's monumental account and documentation of the history of the Vienna Circle is now available in English translation—*The Vienna Circle Studies in the Origins, Development, and Influence of Logical Empiricism* (New York Springer 2001) His book is bound to raise our understanding of the Vienna Circle to an entirely new level The work reveals clearly how logical empiricism in Vienna saw itself from the start as detached from the philosophical tradition and how it was hostile to both Kantian and Hegelian tendencies within academic philosophy The early adherents of the movement were, for the most part, not even professional philosophers but mathematicians, physicists, social theorists, and psychologists They thought of themselves, in the terms of Moritz Schlick's programmatic essay of 1930, as part of a new turn in philosophy ("Die Wende der Philosophie," *Erkenntnis*, I (1930) 4-11) which meant for them, at the same time, a split with established philosophizing Schlick's formulations are, in turn, under the influence of Wittgenstein's *Tractatus*, where the break with the philosophical tradition is thematized from the preface to the last sentence

[10] See my "What Has History to Do with Me? Wittgenstein and Analytic Philosophy," *Inquiry*, XLI (1998) 99-121 I argue there against Peter Hacker's assertion that analytic philosophy originated at the Cam, as well as Michael Dummett's claim that it originated in Austria Instead, I propose a multifactorial account in which the identity of analytic philosophy emerges cumulatively from the work of many This seems to me most evident with respect to the interactions between Frege, Russell, and the early Wittgenstein in which a great deal of the initial agenda of analytic philosophy was defined What entered this agenda were concerns that the three shared (specifically with logical analysis, with structure and logical form) but equally important were their distinctive and individual concerns Frege's focus on epistemological questions, on the relation between the empirical and the a priori, Russell's interest in ontological problems, his pluralism and realism, and finally Wittgenstein's preoccupation with language but also his underlying skepticism For further elaboration of this picture, see my "Macht und Ohnmacht der analytischen Philosophie," in Friedrich Stadler, ed , *Bausteine wissenschaftlicher Weltauffassung* (New York Springer 1997), pp 11-33

Friedman is not unaware of these complexities. In refining his thesis, he writes modestly that his account of Carnap, Heidegger, and Cassirer in terms of the fate of neo-Kantianism is simply meant "to shed as much light as possible on the analytic/continental divide from the point of view of this one particular cluster of ideas" (xiii). But that raises the question why this particular point of view deserves the attention that Friedman gives it. For an answer, we must look at the larger picture he sketches of the history of philosophy since Kant.[11]

The Kantian system represents, on his view, "a remarkable synthesis of virtually the whole of human thought" (145). Post-Kantian idealism, however, abandoned Kant's architectonic in favor of a systematic unity of nature and spirit, of reason and culture. After the demise of idealism, there arose ultimately the neo-Kantians who rejected both Kant's original opposition between sensible and intellectual faculties and the subsequent attempt to establish the metaphysical unity of nature and spirit. The neo-Kantians instead reinterpreted Kant's transcendental logic as an epistemology of the positive sciences of their time. They did so, however, in a Kantian spirit by taking science to mean the whole field of human understanding, that is, both the natural and the human and cultural sciences. But with the destruction of Kant's architectonic, the neo-Kantians could no longer rely on pure logic as the framework for a philosophy of the whole of intellectual and cultural life. Hence the need for Cassirer's "herculean efforts" to construct a comprehensive philosophy of symbolic forms. His attempt to produce such a philosophy proved inadequate, however, and "we are finally left (in the present space of intellectual possibilities, of course) with the fundamental philosophical dilemma presented by Carnap and Heidegger" (156). These two respond in radically different ways to the problems left over by the unaccomplished neo-Kantian synthesis. While Carnap holds on to formal logic as the ideal of universal validity and confines himself to the philosophy of the exact mathematical sciences, Heidegger cuts himself off from logic in pursuit of an understanding of the human condition Friedman concludes

---

[11] Friedman's story is intriguing but highly selective. His account touches barely on the huge influence of empiricism, naturalism, positivism, and Darwinism on the development of nineteenth-century thought. These tendencies were important even in the formation of the neo-Kantian tradition, which was anything but a closed school. The tradition included philosophers attached to idealist and Hegelian modes of thought as well as metaphysical realists and scientific positivists. Hermann von Helmholtz, Friedrich Albert Lange, and Hans Vaihinger who belonged to neither of the two major schools of Neo-Kantianism were as much part of the tradition as Hermann Cohen, Rickert, and Bauch and they played possibly a more direct role in the emergence of analytic philosophy than the latter group.

> If I am not mistaken, it is precisely this dilemma that lies at the heart of the twentieth century opposition between "analytic" and "continental" philosophical traditions, which thus rests, from a purely philosophical point of view, on the systematic cracks in the original Kantian architectonic (156)

In order to overcome the persistent dilemma, Friedman considers it helpful to turn back to Cassirer Not that we can simply appropriate his viewpoint, but Cassirer, he believes, can still define for us an agenda on which analytic philosophy may fruitfully proceed, one which will allow it, moreover, to enter a productive engagement with the concerns of the continental tradition For while Carnap and Heidegger were creating the division between the two traditions, Cassirer was making "an heroic attempt to bridge the ever widening gap between the scientifically oriented approach to philosophy championed by Carnap and the decisive attempt to move philosophy in a quite contrary direction represented by Heidegger" (xii) His work can, for that reason, serve to "provide us with new possibilities and renewed motivation for making a similarly heroic effort for ourselves" (xii) Friedman considers it, in fact, "hard to imagine making progress without increased appreciation for both the strengths and weaknesses" of Cassirer's wide-ranging and deeply synthetic style of philosophical thought (xii) He concludes his monograph accordingly with the words

> Those interested in finally beginning a reconciliation of the analytic and continental traditions can find no better starting point than the rich treasure of ideas, ambitions, and analyses stored in his astonishingly comprehensive body of philosophical work (159)

In order to substantiate his picture of Cassirer as a potential mediator between the analytic and the continental traditions, Friedman turns to an account of the "International University Course" at Davos, Switzerland in the Spring of 1929 in which Cassirer and Heidegger disputed over the interpretation of Kant's critical philosophy The event has become famous in recent years as symbolizing a watershed in the philosophical currents of the time with Cassirer standing for the waning tide of neo-Kantianism, which had dominated German philosophy for half a century, and Heidegger for the new rising waters in philosophy Because of the political turmoil of the following years which forced the Jewish and politically liberal Cassirer to flee Germany and led Heidegger to identify misguidedly with Hitler's regime, the Davos meeting has also come to be seen as a confrontation between Enlightenment and anti-Enlightenment forces in European philosophy. Such a political interpretation of the Davos meeting is questionable, however, because political differences played no role in the Davos discussions, the Enlightenment was not at stake in them, and it is far from clear that Heidegger was even vaguely motivated by political concerns at that time.

Friedman understands that the political interpretation of Davos is problematic, but he, too, emphasizes the difference in the political and social ideals of those involved in it and, while he does not argue the case, he, too, suggests that there is an inherent link between the more technical philosophical views of Carnap, Cassirer, and Heidegger and their political commitments. This will certainly require further argumentation since we know by now that some of Cassirer's fellow neo-Kantians also became National Socialists in 1933 (for example, Bauch), that some logicians and positivists did likewise (for example, Gerhardt Gentzen and Hugo Dingler), and that some existential philosophers were committed democrats (for example, Karl Jaspers). The relations between philosophical views and political commitments are, so it appears, less tidy than we like to imagine.

But the supposedly political side of the Davos debate is, in any case, not of central concern to Friedman. More important to him is the fact that Carnap was in the audience during the three-week event and managed to have conversations with Heidegger (and presumably also with Cassirer). Shortly afterward, he wrote his polemical essay "The Elimination of Metaphysics through the Logical Analysis of Language" (published in 1932[12]) which attacked Heidegger as a typical representative of the metaphysical tradition and set out to expose his assertion that "the nothing nihilates" as characteristically nonsensical. The sharpness of Carnap's words, so Friedman believes, in turn instigated Heidegger to draw a line in the sand between the logic-oriented work of his opponent and his own philosophizing. In a withering blast in his 1935 lectures,[13] he accused Carnap of a nihilistic forgetfulness of the basic question of philosophy (his remarks remaining unpublished until both protagonists were dead). The political currents had by that time taken Carnap and many of his philosophical associates to the English-speaking world, while Heidegger with many other "continental" philosophers were left behind.

---

[12] Translated by A. Pap in A. J. Ayer, ed., *Logical Positivism* (Glencoe, IL: Free Press, 1959), pp. 60-81.

[13] Translated by Ralph Manheim as *An Introduction to Metaphysics* (New Haven: Yale, 1959).

Philosophical disagreement and political history thus conspired to rupture the conversation between the two sides. For Friedman, it is evident that Davos was the point at which "the ensuing split between what we now call the analytic and continental philosophical traditions" was initiated (xi).

The main effect of this retelling of the Davos meeting is, of course, to re-enforce the picture of Cassirer as the man in the middle between Carnap's and Heidegger's extremes. In some ways, he is "extremely close, in fact, to the position of Carnap and the logical positivists" (110). But he is also sympathetic to the phenomenological tradition and "the contemporary *Lebensphilosophie* of Dilthey, Bergson, Scheler, and Simmel" on which Heidegger draws (101). But can Cassirer's work really carry the burden that Friedman assigns to it—that of moving current philosophizing beyond the stultifying analytic/continental divide? Much stands in the way of that. For one thing, contemporary readers of Cassirer may find him heavy going. His writing displays both far-ranging historical scholarship and an overall vagueness in formulation. His high-mindedness in both style and content is at times hard to bear. Friedman records Carnap's note from the Davos meeting that "Cassirer speaks well, but somewhat pastorally" (7). This means, presumably, that Carnap sees him as somewhat of a Protestant pastor and that is, indeed, an apposite characterization of both the content and tone of Cassirer's writings. But the main problem with regarding Cassirer as the protagonist for the next phase of philosophy is not such shortcomings. It is rather what Cassirer stands for. Can we really conceive of the next step in philosophy as a kind of neo-neo-Kantianism? There are distressingly many contemporary philosophers who might respond affirmatively to that question. But before we all rush to the works of Cassirer we may want to ask what kind of mediation between analytic and continental philosophy he can offer us. Friedman praises him and, indeed, the neo-Kantians as a whole for their attempt "to incorporate both scientific and non-scientific modes of thought within a single philosophical framework" (100). And he quotes Cassirer approvingly as writing:

> The natural scientific concept does not deny or negate the object of ethics and aesthetics, even though it cannot construct it with its means, it does not falsify intuition, even though it consciously considers it under a *single* governing point of view and gives prominence to a single determination for it. It is a new teleological order of reality that now appears next to the mere order of magnitudes and in which the individual first obtains its full meaning. (100)

There derives from this an irenic, reconciling, synthesizing conception of the task of philosophy. Its goal is "a single philosophical framework" in which the concerns of both analytic and continental philosophy can be brought together.

Some such reconciliation is, indeed, hoped for by many of those now writing about the divide between analytic and continental philosophy. But one might have a very different view of the current situation and consider the conflict and disagreement between these traditions as both inevitable and desirable. Not that a state of mutual ignorance is good; one certainly hopes for energetic interchanges across the lines of division. But that is not the same as expecting or wanting a reconciliation and even less wanting a reconciliation on the basis of some kind of Kantian schema. If the history of philosophy teaches us anything, it is that the divergence and conflict of schools is a productive force. If Carnap and Heidegger had shaken hands at Davos in the name of Cassirer's synthesis, would twentieth-century philosophy have been better off? What is more, the terms 'analytic' and 'continental' have by now become catch-alls for all kinds of movements and ideas and the disagreements within each tradition have proved conflictual and stimulating as well as resistant to reduction. Given the low chance of a reconciliation between the various factions *within* analytic philosophy and *within* the continental tradition, why should we expect one *between* the analytic and the continental fronts?

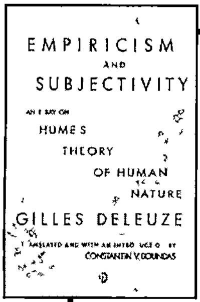
To the neo-Kantian project of finding a single philosophical framework in philosophy we can oppose that of establishing difference, of emphasizing the essential and productive disunity of human knowledge, of critiquing all philosophical attempts at reducing things to a common denominator. Such a philosophy would not be at all systematic and synthetic in the Kantian sense, but topical, case-oriented, and critical in outlook. Wittgenstein, as we know, practiced some such form of thought (and so do some of the postmoderns). To mention these alternatives to Friedman's proposal is not to predict the future course of philosophy. But they remind us how wide the arc of philosophical thinking now is and that the very project of looking for a single philosophical framework is already controversial even before we begin to ask in what terms it should be constructed.

Friedman's discussion remains of interest even if we do not buy into his agenda. It recalls the fact that neo-Kantianism was at one point the dominant force in European philosophy and that both analytic and continental philosophy have grown out of its soil. This is important not, as Friedman imagines, for charting the future of analytic philosophy but for understanding its present character. For nothing is more disconcerting to those committed to other ideals within the analytic tradition (be they naturalistic, empiricist, positivist, scientistic, pragmatist, skeptical, or Wittgensteinian) than the apparently unstoppable encroachment of Kantian modes of thinking. In ethics, political philosophy, and even aesthetics this is only too obvious. It is no less evident in epistemology, in the preoccupation with transcendental arguments and the conditions of the possibility of this or that concept, and it is evident even in the current self-interpretation of much of analytic philosophy. Friedman's observation that neo-Kantianism stood at the beginning of the analytic tradition can explain to us why a creeping Kantianism or neo-Kantianism or neo-neo-Kantianism has proved so attractive. For those who would rather not walk that path, Friedman's monograph is an eye-opening and timely reminder.

HANS SLUGA

University of California/Berkeley

New in **paperback**

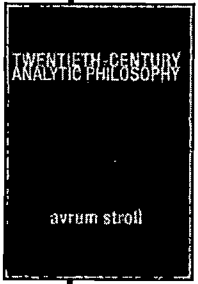


Empiricism and Subjectivity
An Essay on Hume's Theory of Human Nature
Gilles Deleuze Translated by Constantin V. Boundas

Presenting a challenging reading of David Hume's philosophy, this work is invaluable for understanding the progress of Deleuze's thought.

"This book is essential."—David Allison, SUNY, Stony Brook

186 pages • $17.50
EUROPEAN PERSPECTIVES: A SERIES IN SOCIAL THOUGHT AND CULTURAL CRITICISM LAWRENCE D. KRITZMAN, EDITOR



Twentieth-Century Analytic Philosophy
Avrum Stroll

"No other book has canvassed the last century as Stroll has, and with such insight. Written with wit and verve, Stroll's book is informative and eloquent."
—*Journal of the History of Philosophy*

304 pages • $18.50

COLUMBIA UNIVERSITY PRESS
columbia.edu/cu/cup 800-944-8648

NEW BOOKS TRANSLATIONS

KENNEY, DOUGLAS F , and SABINE MENNER-BETTSCHEID, eds and trans *The Recalcitrant Art Diotima's Letters to Holderlin and Related Missives* Albany SUNY, 2000 xiv, 257 p Paper $20 95

KOJÈVE, ALEXANDRE *Outline of a Phenomenology of Right* BRYAN-PAUL FROST, ed BRYAN-PAUL FROST and ROBERT HOWSE, trans Lanham, MD Rowman and Littlefield, 2000 xv, 492 p Cloth $45 00

KRISTEVA, JULIA *Hannah Arendt Life Is a Narrative* FRANK COLLINS, trans Toronto University Press, 2001 vii, 100 p Cloth $18 95, paper $12 00

LAFONT, CRISTINA *Heidegger, Language, and World-Disclosure* GRAHAM HARMAN, trans New York Cambridge, 2000 xxii, 298 p Cloth $59 95

LEFORT, CLAUDE *Writing The Political Test* DAVID AMES CURTIS, trans Durham, NC Duke, 2000 xlii, 312 p Cloth $54 95, paper $18 95

LEVINAS, EMMANUEL *God, Death, and Time* BETTINO BERGO, trans Stanford University Press, 2000 xii, 296 p Cloth $49 50, paper $18 95

LYOTARD, JEAN-FRANÇOIS *The Confession of Augustine.* RICHARD BEARDSWORTH, trans Stanford University Press, 2000 xviii, 111 p Cloth $39 50, paper $12 95

MCGRADE, ARTHUR STEPHEN, JOHN KILCULLEN, and MATHEW KEMPSHALL, eds and trans *The Cambridge Translations of Medieval Philosophical Texts Ethics and Political Philosophy* New York Cambridge, 2000 xii, 664 p Cloth $85 00, paper $29 95

MEYER, MICHEL *Philosophy and the Passions Toward a History of Human Nature* ROBERT F BARSKY, trans University Park Penn State UP, 2000 xxix, 295 p Cloth $65 00, paper $22 50

NANCY, JEAN-LUC *Being Singular Plural.* ROBERT D RICHARDSON and ANNE E O'BYRNE, trans Stanford University Press, 2000 xvi, 207 p Cloth $49 50, paper $17 95

PLATO *The Republic* G R F FERRARI, ed , TOM GRIFFITH, trans New York Cambridge, 2000 xlviii, 382 p Cloth $37 95, paper $10 95

POPPER, KARL, with GIANCARLO BOSETTI *The Lesson of This Century* PATRICK CAMILLER, trans New York Routledge, 2000 vi, 96 p Paper $16 95

ROSENZWEIG, FRANZ *Philosophical and Theological Writings* PAUL W FRANKS and MICHAEL L MORGAN, trans , and eds Indianapolis Hackett, 2000 xii, 165 p Cloth $34 95, paper $12 95

SCHLEGEL, FRIEDRICH *On the Study of Greek Poetry* STUART BARNETT, ed and trans Albany SUNY, 2001 ix, 146 p Cloth $49 50, paper $16 95

SPINOZA, BARUCH *Political Treatise* SAMUEL SHIRLEY, trans Indianapolis Hackett, 2000 xvi, 193 p Cloth $34 95

STENGERS, ISABELLE *The Invention of Modern Science* DANIEL W SMITH, trans Minneapolis Minnesota UP, 2000 184 p Paper $19 95

VOLTAIRE *Treatise on Tolerance and Other Writings* SIMON HARVEY, ed BRIAN MASTERS and SIMON HARVEY, trans New York Cambridge, 2000 xxv, 160 p Cloth $54 95, paper $19 95

THE JOURNAL OF PHILOSOPHY
VOLUME XCVIII, NO 12, DECEMBER 2001

25/1/02

TWO DOGMAS OF DAVIDSONIAN SEMANTICS*

In "Truth and Meaning," Donald Davidson[1] first formulated what was to become known as "Davidson's program " He proposed to elucidate the notion of natural-language meaning in general by showing how to construct a theory of meaning for a particular language, that is, a theory which would allow the interpretation of all the sentences of that language Davidson's basic idea was to exploit a technique that Alfred Tarski invented in his endeavor to show how truth could be defined for a formal language This led to the slogan that a theory of truth for a language could "serve as" a theory of meaning for that language Davidson developed this idea in subsequent years, adding in particular his theory of radical interpretation, which was to explain how a theory of meaning for a particular language (in Davidson's sense) could be empirically confirmed

It is my aim here to re-examine and criticize two doctrines which have become part of the Davidsonian program, but which are not essential to his original idea They are the result, in my view, of a few

---

* I am grateful to Gabriel Segal for his extremely helpful comments on a recent version of this article Parts of it have been presented on various occasions One of these was the XVIth International Symposion of Philosophy, which was held in honor of Donald Davidson in October 1997 at the Instituto de Investigaciones Filosóficas of the UNAM in Mexico City, where I received helpful comments from Davidson, Mario Gómez-Torrente, John McDowell, Rupert Summerton, and other participants Another occasion was Wolfram Hinzen and Hans Rott's workshop on Belief and Meaning at Regensburg in May 2000, where Johannes Brandl, Wlodek Rabinowicz, and others generously helped with comments I would like to thank the audiences of these presentations for their comments and the organizers of these events for inviting me Finally, I would like to thank David Papineau, Michael Potter, Mark Sainsbury, and Mark Textor for discussion and comment

[1] *Synthese*, xvii (1967) 304-23, reprinted in Davidson, *Inquiries into Truth and Interpretation* (New York Oxford, 1984), pp 17-36 Page references will be to the reprinted version (this holds for other cited articles reprinted there)

0022-362X/01/9812/613-35

wrong turns that the development of Davidson's program took during early debates in the 1970s The first of these doctrines is the prima facie absurd view that a theory of meaning for a language does not say what any sentence of that language means More precisely, this is the view that the target theorems of a theory of meaning for a language ought to take the extensional form of material biconditionals of the form '*s* is true if and only if *p*', so that the theorems of a theory of meaning do not *state* what the sentences of the language mean (or what their truth conditions or contents are), but rather "give the meaning" of sentences and allow us to interpret them if we have further information about these theorems I shall call this the *biconditional doctrine* The second doctrine is the view that the concept of truth plays a central explanatory role in Davidsonian theories of meaning for a language I shall call this the *truth doctrine*

I shall argue that the original reasons for adopting these two doctrines are flawed, and that there are, in fact, good reasons for not adopting them Both doctrines are often uncritically accepted and have almost become part of the Davidsonian legacy But, in fact, they are unjustified, and the main insights of Davidson's program do not depend on them That is why, in my title, I have called them "two dogmas of Davidsonian semantics "

I DAVIDSON'S PROGRAM

In "Truth and Meaning," Davidson sets out to describe the form a theory of meaning for a particular language should take if it is to show "how the meanings of sentences depend upon the meanings of words" (*ibid*, pp 17, 23) This pretheoretical adequacy condition is dictated by the simple need to explain the fact that languages can be learned even though they contain an indefinite number of sentences, the meanings of which could not be learned one by one Davidson proceeds by considering the form of the theorems that a theory of meaning would generate He rejects theorems of the form '*s* means *m*', where the replacement for '*m*' is an expression referring to a meaning (he uses a form of the slingshot argument) He also rejects theorems of the form '*s* means that *p*', because

> it is reasonable to expect that in wrestling with the logic of the apparently non-extensional 'means that' we will encounter problems as hard as, or perhaps identical with, the problems our theory is out to solve (*ibid*, p 22)

Davidson's point here is that, if we want a theory that entails theorems of the form '*s* means that *p*', then we need to know something about the logic of the expression 'means that' we need to know which inferences involving this expression are valid But the expression is intensional, and the best available account of the logical properties of intensional contexts involves a notion of meaning or synonymy within intensional contexts, the substitution of synonymous expressions preserves truth[2] But if we are out to explain the notion of meaning in general, or in a particular language, we cannot employ a logic that presupposes the very same notion of meaning

Because of these difficulties, Davidson then looks out for a different, extensional expression that is to fill the gap in theorems of the form ‘*s* *p*’, so that these theorems can be derived in a purely extensional axiomatic system He says that

> the success of our venture depends not on the filling but on what it fills The theory will have done its work if it provides, for every sentence *s* in the language under study, a matching sentence (to replace ‘*p*’) that, in some way yet to be made clear, “gives the meaning” of *s* (*op cit*, p 23)

His suggested filling is ‘is *T* if and only if’, that is, some (initially uninterpreted) predicate ‘is *T*’ combined with the extensional sentential connective ‘if and only if’ The adequacy of a theory of meaning for a language can now be captured by the requirement that the theory “entail all sentences got from” (*op cit*, p 23) the schema ‘*s* is *T* if and only if *p*’ when ‘*s*’ is replaced by a structural description of a sentence of that language, and ‘*p*’ by a translation of that sentence

It then turns out that Tarski’s recursive method of defining a truth predicate for a formal language provides a way of satisfying this adequacy condition on ‘is *T*’ Thus Davidson’s proposal is that at least one good way of explaining how the meanings of the sentences of a given language depend upon the meanings of words is to construct a Tarski-style recursive theory of truth which generates theorems of the form ‘*s* is *T* if and only if *p*’ for every sentence of that language[3] This is Davidson’s basic idea to exploit Tarski’s recursive technique in order to do justice to the compositional requirements of a semantic theory In subsequent works, Davidson developed the details of this approach by (a) showing how some recalcitrant natural-language constructions could be forced into the tight corset of a Tarskian recursive structure, and (b) explaining how a theory of meaning thus understood could be confirmed by empirical data (radical interpretation) Many philosophers and theoretical linguists have since joined Davidson in this effort

---

[2] See Gottlob Frege, “Über Sinn und Bedeutung,” *Zeitschrift fur Philosophie und philosophische Kritik*, c (1891) 25-50, English translation in M Beaney, ed, *The Frege Reader* (Cambridge Blackwell, 1997), see also Rudolf Carnap, *Meaning and Necessity* (Chicago University Press, 1956)

[3] It is important to realize that Davidson never argued that this is the *only* way of constructing such a theory All he says is that so far “we have no other idea how to turn the trick” (*op cit*, p 23)

II THE BICONDITIONAL DOCTRINE

The biconditional doctrine results from an obvious basic problem with Davidson's approach Suppose we have constructed a Davidsonian theory of meaning for a language, that is, a theory that entails a theorem of the form '*s* is *T* if and only if *p*' for each sentence *s* of the language, such that what replaces '*p*' in each theorem is a translation of *s* These theorems are, in common parlance, "interpretive" ("give/show the truth condition of a sentence"), but they do not *state* what the mentioned sentence means (or what its truth conditions are) They are just material biconditionals all that is required for the truth of a material biconditional is that both sides flanking the biconditional have the same truth value If I know that a given sentence is true if and only if snow is white, I do not thereby know that this sentence means that snow is white, nor do I thereby know that its "truth condition" is that snow is white An easy way to see this is to consider the two sentences

(T1) 'Snow is white' is *T* if and only if snow is white
(T2) 'Snow is white' is *T* if and only if grass is green

If we wanted to say that by knowing (T1), I know that 'snow is white' means that (or has the truth condition that) snow is white, then we would have to say also that by knowing (T2), I know that 'snow is white' means that (or has the truth condition that) grass is green

Davidson sets himself this problem in "Truth and Meaning," and his response is to say that in order for (T1) to give me information on what 'snow is white' means, I need to know in addition that (T1) has the status of a natural law, that is, is derivable from a truth theory that has been empirically confirmed in the right way (that is, through a process of radical interpretation) and is maximally simple[4] (T2) is not so derivable—to derive it, one would need further, nonsemantic information on whether the proposition that 'snow is white' is true and the proposition that grass is green have the same truth value

This, however, does not remove all problems Consider

(T3) 'Snow is white' is *T* if and only if snow is white and either grass is green or grass is not green

---

[4] See *op cit*, pp xiv, xviii, 26, and "Reply to Foster," in *Inquiries into Truth and Interpretation*, p 174

Just as (T2), (T3) is not interpretive But unlike (T2), (T3) *is* derivable from a truth theory confirmed by a Davidsonian process of radical interpretation At least it is thus derivable if (T1) is, because it is logically equivalent to (T1) [5]

Davidsonians normally respond to this new difficulty by introducing the notion of a "canonical T-theorem " A T-theorem is a metalanguage sentence of the form '*s* is *T* if and only if *p*' where '*s*' is a description of an object-language sentence and '*p*' is a metalanguage sentence that does not mention any object-language expressions A canonical T-theorem, now, is a T-theorem that can be derived following a specified (canonical) procedure Roughly, this procedure involves applying the semantic axioms concerning the syntactic constituents of *s* in an order that inverts the order in which *s* was constructed from its constituents, and then arriving at a T-theorem by repeated application of the rule of substitution of material equivalents [6] All canonical theorems are interpretive and do not suffer from (T3)'s problem Thus, if one knows that a theorem has been derived in the canonical way, one thereby knows that the sentence mentioned on one side of it is interpreted by the sentence used on the other side

A modified response is given by Richard Larson and Gabriel Segal (*op cit*, §2 2 1-2, cf Segal), who argue that a theory of meaning is supposed to model the knowledge which explains speakers' linguistic behavior, that is, it models the "semantic module " On their view, such a semantic theory consists of a set of semantic axioms and a set of rules of inference ("production rules")—rules that permit fewer inferences than classical logic These inference rules are designed to permit only the derivation of interpretive T-theorems This is an improvement on the standard response, because it avoids the detour of first formulating a theory with general logical inference rules, and then restricting the use of these rules by introducing the notion of a canonical proof If a semantic theory is to model the information contained in the semantic module, such a gratuitous detour ought to be avoided (*op cit*, p 559, footnote 14)

---

[5] This difficulty was originally raised by John Foster in "Meaning and Truth Theory," in Gareth Evans and John McDowell, eds , *Truth and Meaning* (New York Oxford, 1976), pp 1-32 For excellent discussion of Davidson's way around the difficulty, see Gabriel Segal, "How a Truth Theory Can Do Duty as a Theory of Meaning," in U Zeglen, ed , *Donald Davidson Truth, Meaning, and Knowledge* (New York Routledge, 1999), pp 48-58 Segal thinks that (T3) could be ruled out on the grounds that it is not part of a maximally *simple* theory

[6] Martin Davies spells out such a canonical proof procedure in his *Meaning, Quantification, Necessity* (New York Routledge, 1981), p 33 See also Christopher Peacocke's "Truth Definitions and Actual Languages," in Evans and McDowell, pp 162-88, and Richard Larson and Segal, *Knowledge of Meaning* (Cambridge MIT, 1995)

Both the standard and the modified response to the problem leave Davidsonians with the awkward biconditional doctrine the theorems of a theory of meaning are material biconditionals which do not *state* what the sentences mentioned mean They do, however, "give" the meaning, or truth condition, of the sentence, and additional information of what constitutes a canonical proof allows one to use such a theory for interpretive purposes In Larson and Segal's formulation humans have a semantic module which can be modeled as a T-theory, and humans treat the theorems that can be generated by the theory *as* interpretive (*op cit*, p 39, Segal, p 55)

It is surprising that the cumbersome biconditional doctrine has not received more critical attention The view that a semantic theory, or a semantic module, does not, on its own, provide information on what sentences mean should have been highly suspect Larson and Segal, who are the only Davidsonians who take the problem seriously, improve upon the standard version of the doctrine But their claim that, by allowing only a restricted set of inference rules, one can avoid uninterpretive theorems should have led to further reflection if the only T-theorems derivable in a semantic theory are interpretative ones, then it should have been possible to modify the theory in such a way that it generates genuinely meaning-specifying theorems of the form '*s* means that *p*' (or '*s*'s content is that *p*' or '*s*'s truth condition is that *p*')

It is easy to see that the unattractive biconditional doctrine is unnecessary, if one considers Davidson's original reason for introducing target theorems of the form '*s* is *T* if and only if *p*' Davidson's reason, as mentioned above, was the intensionality of the expression 'means that' He thought that the theorems of a theory of meaning could not take the form '*s* means that *p*' because in order to derive such theorems, he would need to know intensional logic, something he thought presupposed the notion of meaning But both the standard version and the modified version of the biconditional doctrine provide an easy solution to this problem of deriving intensional theorems, or so I shall argue in the next section

III HOW TO DERIVE INTENSIONAL THEOREMS

Let me explore in more detail how genuinely meaning-specifying theorems of the form '*s* means that *p*' could be derived in a semantic theory of Tarski-Davidsonian cut

Consider the standard version of the biconditional doctrine first It says that there is a canonical procedure following which one can derive, from the semantic axioms of the truth theory, all and only interpretive T-theorems It would seem that this provides us with all we need to know about the intensional logic of 'means that' Can we not simply add an inference rule that permits one to move from

(P) 's is T if and only if p' is a canonical T-theorem

to

(C) s means that p

Let me first make a general observation about the form of inference rules Any inference rule for a theory of meaning M would be formulated not in the language L_M of M itself but in a metatheoretic language It would take the form of an inference schema with schematic letters ranging over expressions of L_M (not the object language) The schema then indicates that certain inferential moves from sentences of L_M to sentences of L_M are permitted This is not unusual the inference rules of any theory can be formulated only metatheoretically

The suggested inference schema is not a rule of this sort it permits inferences from sentences of the form of (P) to sentences of the form of (C) The former, however, are not sentences of L_M, but metatheoretic This is because instances of (P) mention L_M sentences and ascribe to them the metatheoretically defined property of being a canonical T-theorem

We need a proper metatheoretic rule that permits a move within L_M from a T-theorem to its meaning-specifying counterpart if the T-theorem is canonically derived The following formulation uses an adapted version of a standard notation for schematically stating inference rules

(R)

(canonical derivation)

$$\frac{s \text{ is } T \text{ if and only if } p}{s \text{ means that } p}$$

Schema (R) shows that one may derive an L_M sentence 's means that p'[7] if one has previously been able to derive an L_M sentence 's is T if and only if p' in the canonical way The three dots, together with the specification in brackets, give a metatheoretic instruction they indicate that '*s* is *T* if and only if *p*' needs to have been canonically derived if the move to '*s* means that *p*' is to be legitimate

---

[7] For convenience, I use ordinary quotes where corners would be appropriate

This form of stating a rule of inference may appear unorthodox and therefore arouse suspicion But a little reflection will show that it is not unorthodox Consider, for example, a typical schematic formulation of the rule of conditional proof for the propositional calculus

$$\text{(CP)}\quad \begin{array}{c} [p] \\ \\ \\ \\ q \\ \hline p \supset q \end{array}$$

Consider the role of the metatheoretic instruction here the three dots between the bracketed '*p*' and '*q*' indicate that, if a formula *q* has been proved on the assumption that *p*, then one may infer '$p \supset q$' Standard formulations of other rules of inference, such as reductio ad absurdum and constructive dilemma, involve even more complex metatheoretic instructions[8] Thus, if these are legitimate formulations of inference rules (which I take it they are) then so is (R)

But if (R) is a legitimate form of characterizing a rule of inference, then all we need to make it work is a canonical procedure following which one can derive all and only interpretive T-theorems As we saw, the standard move of Davidsonians in the face of Foster-type problems is to invoke the existence of precisely such a canonical procedure[9] Thus, there is no reason why these Davidsonians could not introduce a rule like (R), thereby generating intensional, meaning-stating theorems and giving up the biconditional doctrine

Now consider Larson and Segal's modified version of the biconditional doctrine They claim that the inference rules "contained in the semantic module" allow the derivation of all *and only* interpretive T-theorems If this is true, then there may be an even more direct way of deriving meaning-stating theorems just add an inference rule that allows the move from any T-theorem to its meaning-specifying counterpart Schematically

---

[8] Compare, for example, Wilfred Hodges, "Elementary Predicate Logic," in D Gabbay and F Guenthner, eds , *Handbook of Philosophical Logic*, Volume I (Dordrecht Reidel, 1983), pp 1-131, especially pp 29-30

[9] See page 617 above and footnote 6

$$(\text{R}^*)\quad \frac{s \text{ is } T \text{ if and only if } p}{s \text{ means that } p}$$

Since a Larson-and-Segal-style semantic module employs inference rules that do not allow the derivation of uninterpretive T-theorems, we need no metatheoretic instruction that restricts this move to cases where the premise has been canonically derived (*op. cit.*, p. 40, footnote 15).

This is too simple, however. Not every properly derived L_M sentence of the form '*s* is T if and only if *p*' is a T-theorem, that is, a theorem in which the right hand side '*p*' is replaced by an L_M sentence which does not mention any object-language expressions. As a consequence, (R*) allows too much. For example, the theory will allow the derivation of sentences of the form "'*s* & *r*' is true if and only if *s* is true and *r* is true." But "'*s* & *r*' means that *s* is true and *r* is true" is false. Our inference rule should only be applicable to T-theorems. This needs to be included in the metatheoretic description of the premise in (R*). So the following might be a correct formulation of the rule for Larson and Segal's theory:

$$(\text{R}^{**})\quad \frac{\begin{array}{l} s \text{ is } T \text{ if and only if p} \\ \text{(where `}s\text{' is a description of an object-language sentence and `}p\text{'} \\ \text{is an } L_M \text{ sentence that does not mention object-language expressions)} \end{array}}{s \text{ means that } p}$$

Thus, both on the standard view and on Larson and Segal's, there is a way in which a theory of meaning can have intensional, meaning-specifying theorems.[10]

---

[10] Some theorists might require that all inference rules be encapsulated in the formal definition of a derivability notion. Adding a rule of inference, on this view, must take the form of adding a clause to the definition of 'derivable' or 'theorem'. This requirement is unproblematic, at least for my proposed modification of Larson and Segal's theory. Since it is already the case that only interpretive T-theorems are derivable in their theory, and since the definition of 'T-theorem' (as well as the corresponding metatheoretical instruction in (R**)) is in purely syntactic terms, (R**)'s work can be done by a clause like the following:

(R***) If '*s* is *T* if and only if *p*' is a T-theorem, then '*s* means that *p*' is also a theorem.

Things are less straightforward in the case of the proposal, made above, to add (R) to the T-theories of more standard Davidsonians (that is, those who operate with the notion of a canonical theorem). Unlike Larson and Segal, these theorists apply general deductive rules to their axioms, which then generate noninterpretive T-theorems. In order to do (R)'s work in a formal definition of derivability it may be necessary to start from scratch with a notion of derivability that does not allow

All this, however, should not be taken to suggest that we can do without T-theorems or that we can replace '— is true if and only if ' early-on, or throughout, with '— means that '. In the derivation of a meaning specification of a sentence, application of (R) or (R**) is only the last step. The real work is done previously by a derivation of a T-theorem.

An example may illustrate why: consider the derivation of a T-theorem for some conjunctive sentence 's & r'. One would first use the axiom for '&' to derive

(i) 's & r' is T if and only if s is T and r is T

then one would use independently derived theorems of the form

(ii) s is T if and only if p
(iii) r is T if and only if q

to derive something of the form

(iv) 's & r' is T if and only if p and q

using a rule of substitution of equivalents. No meaning-specifying theorems could play the role of (ii) and (iii) in this derivation. Of course, (R) could be applied to (ii) and (iii) directly, and would then yield the correct meaning specification for s and r. But in order to derive the meaning specification for 's & r', we need the original biconditional version of (ii) and (iii).

My proposal is therefore not intended as an objection to Davidson's view that "we have no other idea how to turn the trick" (*op. cit.*, p. 23) of formulating a theory that allows one to generate pairings of object-language sentences with their metalanguage translations from information about simple sentence constituents. Tarski's machinery has an indispensable role in the theory (until we find a different idea how to turn the trick). What I object to is the unreflected doctrine that a theory of meaning cannot say what sentences mean because we lack information about the logic of the intensional phrase 'means that'.

Thus, we know enough about the intensional logic of 'means that' in order safely to derive theorems of the form 's means that p' as a final step. We *can* have a meaning theory that states what sentences mean. The cumbersome biconditional doctrine is an unmotivated

---

the derivation of noninterpretive T-theorems in the first place (that is, basically Larson and Segal's strategy).

dogma We *can* after all utilize Tarski-style recursive machinery and still derive, in a second step, what sentences mean [11]

IV DAVIDSON'S PROGRAM BEFORE THE TRUTH DOCTRINE

Once one has taken the step of recognizing that the biconditional doctrine is a dogma, it becomes easier to take an instrumentalist view of the role of the predicate 'is *T*' in some theorems of a theory of meaning it enables the recursive machinery to generate interpretive T-theorems—and ultimately theorems of the form '*s* means that *p*' The important function of the predicate is that it allows us to generate theorems that pair object-language sentences with their metalanguage interpretations

At the time of writing "Truth and Meaning," Davidson was promoting this view himself, as is shown by the remark I quoted above

> the success of our venture depends not on the filling but on what it fills The theory will have done its work if it provides, for every sentence *s* in the language under study, a matching sentence (to replace '*p*') that, in some way yet to be made clear, "gives the meaning" of *s* (*op cit*, p 23)

Davidson continues in this vein later in the paper, when he discusses the problem of evaluative sentences Evaluative sentences would constitute an obvious problem for Davidsonian meaning theories, if the predicate involved in the T-theorems were thought of as expressing the notion of truth For it is controversial whether evaluative sentences (or the contents expressed by utterances of them) can be evaluated in terms of truth at all But Davidson brushes any such worries aside, asserting, once again, that all that counts is whether the theory can generate the right theorems from its axioms, implying that it does not matter whether the predicate involved expresses the notion of truth

> If we suppose questions of logical grammar settled, sentences like 'Bardot is good' raise no special problems for a truth definition The deep differences between descriptive and evaluative (emotive, expressive, and so on) terms do not show here Even if we hold there is some important sense in which moral and evaluative sentences do not have a truth value (for example because they cannot be verified), we ought not to boggle at "'Bardot is good' is true if and only if Bardot is good", in a theory of truth, this consequence should follow with the rest, keeping track, as must be done, of the semantic location of such sentences in the language as a whole—of their relation to generalizations, their role in such

---

[11] This conclusion does not affect Davidson's reasons for rejecting theorems of the form '*s* means *p*' where '*p*' *refers* to a meaning (that is, his slingshot argument)

> compound sentences as 'Bardot is good and Bardot is foolish', and so on What is special to evaluative words is simply not touched the mystery is transferred from the word 'good' in the object language to its translation in the metalanguage (*op cit*, p 31)

These remarks show that Davidson did not object to the T-concept invoked in the T-theorems being distinct from the concept of truth—or at least from truth in any sense in which it cannot be applied to evaluative sentences As long as a theory of meaning delivers the needed T-theorems which permit us to interpret speakers, it does not matter what exactly we mean by 'true' in the theory of meaning Thus, at the time of writing "Truth and Meaning," Davidson did not subscribe to the truth doctrine—the view that the notion of truth plays a key explanatory role in theories of meaning for particular natural languages

V THE EMERGENCE OF THE TRUTH DOCTRINE

Davidson changed his view on this when he developed the theory of radical interpretation [12] It was considerations about the explanatory aims of Tarski's project as opposed to the explanatory aims of his own meaning theories which prompted this change of view and gave rise to the truth doctrine Davidson expressed the new view in many of his works after 1973, for example, in the introduction to *Inquiries into Truth and Interpretation*

> [W]hile Tarski intended to analyze the concept of truth by appealing (in convention T) to the concept of meaning (in the guise of sameness of meaning, or translation), I have the reverse in mind I considered truth to be the central primitive concept, and hoped, by detailing truth's structure, to get at meaning [13]

In "Radical Interpretation" [14]

> [A]ssuming translation, Tarski was able to define truth, the present idea is to take truth as basic and to extract an account of translation or interpretation (*ibid*, p 134)

And in "Belief and the Basis of Meaning" [15]

---

[12] See his remarks in the introduction to *Inquiries into Truth and Interpretation*, pp xiv-xv, and "The Structure and Content of Truth," this JOURNAL, LXXXVII, 6 (June 1990) 279-328, especially p 286, footnote 20

[13] *Inquiries into Truth and Interpretation*, p xiv

[14] *Dialectica*, XXVII (1973) 313-28, reprinted in *Inquiries into Truth and Interpretation*, pp 125-39

[15] *Synthese*, XXVIII (1974) 309-23, reprinted in *Inquiries into Truth and Interpretation*, pp 141-54

> Our outlook inverts Tarski's we want to achieve an understanding of meaning or translation by assuming a prior grasp of the concept of truth (*ibid*, p 150)

And more recently in "The Structure and Content of Truth"

> The theory is correct because it yields correct T-sentences, its correctness is tested against our grasp of the concept of truth as applied to sentences
>
> [I]t is our grasp of [the concept of truth] that permits us to make sense of the question whether a theory of truth for a language is correct (*op cit*, p 300) [16]

Let me spell out some of the background of these remarks In "The Concept of Truth in Formalized Languages," Tarski formulates an adequacy condition for definitions of truth for a language This condition requires that an adequate definition of truth for a language *L* must entail for each sentence of *L* a theorem of the form '*s* is true in *L* if and only if *p*', where '*s*' is a "structural description" of that sentence and '*p*' is the translation of that sentence in the language in which the definition is stated Thus, in order to apply this criterion for the adequacy of a definition of 'true in *L*', one needs to understand what it is for a sentence to be a translation of (or synonymous with) another sentence Now, Davidson proposes to construct a structurally similar theory, also entailing theorems of the mentioned form, but with the aim of "giving" the meaning of each sentence of the language under discussion In order to decide whether such a theory is adequate, Davidson cannot, obviously, presuppose knowledge of which sentences of the metalanguage are translations of which sentences of the object language, as this is what the theory is supposed to yield knowledge *of* So Davidson needs to apply a different criterion of adequacy for his theories of truth (which are to serve as theories of meaning) How could it be decided whether a Davidsonian theory of truth is adequate?

---

[16] This quote may seem to permit the following reading we need to grasp the concept of truth because we need to assess whether the theory, its theorems, are correct, that is, whether they are true This, however, seems to me to be the wrong reading, for one of the most important lessons Tarski draws in "The Concept of Truth in Formalized Languages"—in *Logic, Semantics, Metamathematics* (New York Oxford, 1956), pp 152-278—is that we must distinguish 'true' as a predicate of the object language and 'true' as a predicate of the metalanguage The concept of truth whose grasp is required according to the current reading would correspond to a predicate in the meta-metalanguage, as it gets applied to metalanguage sentences Thus it would be distinct from the truth notion allegedly required to understand the T-theory itself and its theorems Moreover, if this were the sense in which an understanding of truth was required, then this requirement would not be a special feature of meaning theories, but rather a requirement of *any* theory

Davidson does, in fact, have a very good answer to this question It is given by his methodology of radical interpretation But he also uses the following line of thought to convince himself that he needs to presuppose the notion of truth as a basic explanatory concept assuming what it is for one sentence to translate another, "Tarski was able to define truth,"[17] so if one wants to explain meaning (translation) using the very same theory, one obviously needs to be in possession of the notion of truth already In Paul Horwich's[18] words

> we would be faced with something like a single equation and two unknowns knowledge of the truth conditions of a sentence cannot simultaneously constitute *both* our knowledge of its meaning *and* our grasp of truth for the sentence (*ibid*, p 68)

Davidson concludes that the notion of truth must play a central explanatory role in the construction of theories of meaning

The starting point of this line of thought is correct But its conclusion is not It is correct that Davidson needs a new criterion of adequacy, different from Tarski's He cannot check whether all the theorems of a theory of meaning for a language are correct by checking whether their right-hand sides translate the sentence mentioned on their left-hand sides But he does not need to check whether the theorems are correct by checking *immediately* whether their right-hand sides give sufficient and necessary conditions for the truth of the sentences mentioned on their left-hand sides If he needed to be able to know necessary and sufficient conditions for the truth of all the sentences *immediately*, then the project of interpreting an unknown language would be hopeless [19] Instead, what Davidson can do (and what he, in effect, proposes to do in his methodology of radical interpretation) is to check whether the theorems, if taken as interpretive, allow one to make good sense of the linguistic behavior of the speakers of the language

---

[17] "Radical Interpretation," p 134 This is not in fact an accurate description of what Tarski does in his article, as he did not define truth, but only defined truth in a particular formal language, as Davidson never gets tired of pointing out in later works—for example, in "The Structure and Content of Truth," and in "The Folly of Trying to Define Truth," this JOURNAL, XCIII, 6 (June 1996) 263-78

[18] *Truth* (New York Oxford, 1998, second edition)

[19] And in any case, if *that* were required, then it would not be *the* notion of truth that is presupposed, but the notion of truth in the language under discussion

John McDowell,[20] in his reconstruction of the original Davidsonian project, exploits just this idea A theory of meaning for a language ought ultimately to provide us with information that would be sufficient to interpret speakers of that language correctly Now, a Davidsonian T-theory can take us part of the way It can help us assign to every sentence a propositional content But in addition to such a theory of content, we need a theory of illocutionary force These two elements together form what McDowell calls a "bipartite" theory of meaning If one knows such a theory for a language (and if one has sufficient time and patience) then it allows one to redescribe speakers' phonetic acts, that is, acts of emitting certain sequences of sounds, as propositional acts, for example, acts of asserting that, or asking whether such-and-such According to McDowell, the acceptability of such a theory, as an empirical theory about the language's speakers, is measured by the extent to which these redescriptions allow us to make sense of the linguistic and nonlinguistic behavior of the speakers (*ibid*, pp 44-45) The working of such a theory, and the process of empirically confirming it, in no way depends on the interpretation of the predicate 'is *T*' that we are employing in our interpretive T-theorems and in generating them Nothing prevents us from regarding 'is *T*' as a theoretical notion which is implicitly defined by the theory Any (perhaps partial) coextensiveness between 'is *T*' and our ordinary notion of truth is something we discover afterward, it is not something we need to assume before we start the project In McDowell's own words

> The thesis should be not that [meaning] is what a theory of truth is a theory of, but rather that truth is what a theory of [meaning] is a theory of (*ibid*, p 47)

On McDowell's view, as long as we can make out a legitimate empirical methodology for constructing bipartite meaning theories, nothing obliges us to think of the T-notion employed in the theory of content as the notion of truth

Again, rejecting the biconditional dogma puts one into a better position to see this If we take a meaning theory for a language to yield predictions (theorems) of the form '*s* means that *p*', and not of the form '*s* is true if and only if *p*', then Davidson's motivation for the truth doctrine vanishes completely Obviously, a radical interpreter is not required to know *immediately* whether the meaning specifying theorems delivered by her theory of content are correct She will test their correctness via further predictions these theorems allow us to make about speakers Now, what further predictions these are will depend on how a theory of meaning *interacts* (to use a McDowellian expression) with other theories, that is, theories that predict which sentences speakers will utter given that they have a certain meaning, or in other words theories that explain why speakers utter sentences with a certain meaning

---

[20] "Truth Conditions, Bivalence, and Verificationism," in Evans and McDowell, pp 42-66, especially pp 44-45

Since the truth dogma is deeply entrenched, these abstract and general considerations are likely to meet with scepticism But it should be agreed even by skeptics that the method by which a Davidsonian theory of meaning is to be tested empirically should be our ultimate touchstone for the truth doctrine if the methodology of radical interpretation requires explanatory use of the notion of truth, then the truth doctrine is justified If it does not, then the truth doctrine can be thrown onto the scrapheap of unjustified dogmas together with the biconditional dogma In the next section, I shall therefore look in more detail at a methodology of radical interpretation in which the notion of truth makes no appearance whatsoever[21]

VI RADICAL INTERPRETATION WITHOUT THE NOTION OF TRUTH

How can a Davidsonian theory of meaning for a particular natural language be empirically tested? In order to do this, we need to know more about the *observable* consequences of such a theory What observable consequences could a theory have that states what the sentences of a language mean? Intuitively, if the theory is correct, speakers of the language will use certain sentences under certain conditions For example, it would seem that, if some sentence meant that snow is white, then speakers would have a tendency to utter that sentence when they wish to get across that snow is white If they do in fact have this tendency, then this confirms the theory, if they do not, it disconfirms the theory

This commonsensical strategy is in principle correct But we need to add a bit of theory before we can make some such strategy work We need to clarify the connection between facts of meaning, as specified in the theorems of the modified Davidsonian theory, and the behavior of language users these facts should lead us to expect

---

[21] Even though I follow McDowell in his assessment of the role of the concept of truth in these theories, I depart from him in the details of my account of radical interpretation, in particular in my treatment of illocutionary force (communicative function)

By assigning to each sentence a meaning, or truth condition, via a Davidsonian theory, one has not yet captured all meaning features relevant for communication. Sentences can also be classified as having various communicative functions. Some sentences serve to make assertions, others to ask questions or to issue commands. Combining a Davidsonian theory with a theory of communicative function, one can say that a Davidsonian theory specifies what the *content* of each sentence is, or what *proposition* it expresses, while the theory of communicative function specifies for the performance of which communicative act in relation to this content the sentence serves. For example, uttering the English sentence 'Sam smokes', one *asserts that* Sam smokes, while uttering the sentence 'Does Sam smoke?', one *asks whether* Sam smokes, performing a different communicative act on the same content.

Adding an assignment of communicative functions to a Davidsonian theory of content brings us closer to being able empirically to verify the now combined theory. The combined theory predicts that speakers perform certain actions of assertion, question, and so on when they utter sentences. But how do we test empirically whether a speaker is really performing such an action? We need a detailed account of the various linguistic acts, so that we can test the predictions the theory yields against general psychological assumptions concerning action.

The most promising accounts treat communicative acts as actions in intentional conformity with conventional rules and ultimately motivated by communicative aims. Philosophers like Ludwig Wittgenstein, J. L. Austin, H. P. Grice, John Searle, David Lewis, and Robert Stalnaker have pioneered this type of approach. Speakers aim to influence the beliefs of their audiences and audiences aim to acquire new information. Speakers know that audiences generally respond to utterances in conformity with the rules, and audiences know that speakers generally make utterances in conformity with the rules. This knowledge allows speakers and audiences to further their communicative aims by acting in accordance with the rules. There are characteristic rules for each kind of linguistic act (assertion, question, and so on).

Up to this point, most theorists of speech acts are in agreement. But they diverge considerably when it comes to the nature of the rules and the nature of the conventionality involved. I shall briefly sketch two approaches that seem to me to be particularly promising and comment on their suitability for a methodology of radical interpretation. For simplicity, I shall discuss only assertion.

The first approach is broadly along the lines of Lewis's[22] game-theoretical view of linguistic conventions and explains communicative acts directly in terms of belief-desire psychology. He defines a convention as a certain type of regularity in the behavior of a population of agents who face a recurring coordination problem. A coordination problem is a game-theoretical situation in which there are several equilibria on one of which the agents need to coincide. A convention is a solution to such a recurring problem: agents conform to a certain regularity of behavior because they expect the others to conform to the same regularity and it is in their interest to conform if the others conform. Thus conventional behavior can be explained directly in terms of a simple belief-desire psychology. In the case of linguistic conventions, in particular conventions regarding assertoric sentences, the relevant regularity might require agents (i) to utter a sentence assertoric of a content *p* only if *p*, and (ii) to respond to utterances of such a sentence by coming to believe that *p*.

This rule is only a first shot. It wrongly assumes that speakers who conform to linguistic convention are always truthful and sincere, and that audiences always believe what they are told. Clearly, speakers often inadvertently assert a content that *p* even though it is false that *p*. Moreover, speakers often deliberately assert *p* even though they do not believe *p* themselves. Similarly, audiences often fail to believe contents that have been asserted, or even fail to believe that the asserter believes what she has asserted. But let us assume that we can solve this problem by modifying Lewis's account in the following way.[23] The appropriate speaker regularity is that of asserting *p* only if one either (i) believes that *p*, or (ii) wants to give the impression that one believes that *p*, or (iii) wants to give the impression that one wants to give the impression that one believes that *p*, or …. And the corresponding audience regularity is that of responding to assertions of *p* by either (i) coming to believe that the utterer believes that *p*, or (ii) coming to believe that the utterer wants to give the impression that he believes that *p*, or (iii) coming to believe that the utterer wants to give the impression that he wants to give the impression that he believes that *p*. As before, these regularities are conventions in Lewis's game-theoretical sense.

---

[22] *Convention* (Oxford: Blackwell, 1969), and "Languages and Language," in Keith Gunderson, ed., *Minnesota Studies in the Philosophy of Science, Volume VII* (Minneapolis: Minnesota UP, 1975), pp. 3-35, reprinted in Lewis, *Philosophical Papers, Volume I* (New York: Oxford, 1983), pp. 163-88.

[23] As I argue in my "Lewis, Language, Lust and Lies," *Inquiry*, XLI (1998): 301-15.

We would then have a way of testing our combined theory against general psychological assumptions Suppose the theory says that some sentence *s* is assertoric and expresses the proposition that *p* If the theory is correct, then speakers will expect that audiences of *s* come to believe that the utterer either believes that *p* or wants to give the impression that he believes that *p* or   If speakers' communicative aims together with these expectations provide a good explanation of utterances of *s*, then the theory is confirmed The same goes, ceteris paribus, for audience behavior

The advantage of such an account is that it offers an immediate integration of the combined theory into a general belief-desire psychology Speakers have mutual knowledge of their conformity to certain regularities, and this, together with their communicative desires, motivates them, via ordinary instrumental reasoning, to engage in linguistic action [24] On this model, a methodology of radical interpretation makes no explanatory use of the notion of truth At least it does not so long as the notions of belief, desire, and their contents are independent of the notion of truth But there is no prima facie reason to believe that these notions depend on the notion of truth—except perhaps Davidson's own reason for the truth doctrine, which I have discredited above

The second approach I want to discuss might be called a "conversational approach to communicative action " It differs from the Lewisian approach in two ways First, instead of Lewis's simple regularities concerning individual utterances of sentences, the conversational account states rules that specify the role of assertions in entire conversations, that is, in interconnected series of assertions by several agents Second, the conversational approach makes these rules a matter of social norms, while Lewis's game-theoretical approach denies social norms or sanctions any role in linguistic convention As an example for a conversational account, I shall here use the account of assertion proposed by Robert Brandom,[25] who argues that there is a system of social norms and rules that governs our linguistic interactions Within this system, assertion has a central role an assertion that *p* counts (i) as an undertaking to justify *p* if challenged to do so, and (ii) as issuing a license to use *p* as a premise Quite obviously, these rules rely on the normative vocabulary of social duties and licenses, which are ultimately explained in terms of notions such as authority and sanction Thus a conversational account of assertion does not immediately yield reductive explanations of utterances in terms of the beliefs and desires of the utterer All the same, the account provides an empirical test for bipartite theories of meaning for particular languages If a general theory of action together with a candidate theory of meaning—now combined with a theory of communicative acts interpreted as social linguistic acts, can explain speakers' behavior, then this confirms the candidate theory

---

[24] A disadvantage of such an account is that the knowledge it attributes to speakers is at best implicit knowledge Speakers do not go explicitly through the kind of instrumental reasoning this account suggests The status of a combined meaning theory as a psychological hypothesis would therefore be unclear Compare Stephen Laurence, "A Chomskian Alternative to Convention-Based Semantics," *Mind*, cv (1995) 269-301

[25] "Asserting," *Noûs*, xvii (1983) 637-50, and *Making it Explicit* (Cambridge Harvard, 1994) Another example would have been Stalnaker's pragmatic theory —for example, in "Assertion," reprinted in *Context and Content* (New York Oxford, 1999), pp 78-95, which has been further developed by Lewis, "Scorekeeping in a Language-Game," *Journal of Philosophical Logic*, viii (1979) 339-59 But, since Stalnaker always speaks of contents as truth conditions (which, in turn, he takes to be

Again, if a conversational account of communicative acts is used in our methodology of empirically testing a combined theory of meaning, there is no reason to believe that the notion of truth plays an explanatory role [26] Thus, two promising theories of the communicative acts are at our disposal for use in a truthless methodology of radical interpretation Nothing therefore prevents us from viewing the predicate 'is *T*' as contextually defined by the theory in which it occurs There is no need to assume that the predicate expresses some pretheoretically familiar notion of truth, or that this pretheoretical grasp is required to endow the theory with explanatory power [27] This

---

sets of possibilities), I prefer to use, for current purposes, an account that does not in any way appear to make explanatory use of the notion of truth

[26] The Brandom account fares slightly better than the Lewisian as far as the needed independent account of the contents of belief and desire is concerned There is no danger that the use of the notion of content of belief and desire reimports reference to the notion of truth, for Brandom already provides an independent account of this notion the content of an assertion is constituted by its (material) inferential relations with others, that is, by what would count as a justification of the assertion and for what it would count as justification To count as justification, of course, is another notion within the theory of social action, to be explicated, ultimately, in terms of authority and sanction

[27] There are reasons to believe that 'is *T*' as contextually defined by a theory of meaning for a language *L* will be coextensive with 'is true' in those areas where both can be applied See McDowell, "Meaning, Communication and Knowledge," in Zak Van Straaten, ed , *Philosophical Subjects Essays Presented to P F Strawson* (New York Oxford, 1980), pp 117-39, here p 121, and Wiggins, "What Would Be a Substantial Theory of Truth?" in Van Straaten, ed , pp 201, 203-04, and "Meaning, Truth-Conditions, Proposition Frege's Doctrine of Sense Retrieved, Resumed and Redeployed in the Light of Certain Recent Criticism," *Dialectica*, XLVI (1992) 61-90 But this does not show that 'is *T*' and 'is true' express the same concepts As McDowell emphasises, the coextensiveness of the two predicates, where the ranges over which they are defined overlap, should come as a *discovery*

confirms the conclusion of my earlier argument (in section v) that the truth doctrine is a dogma.

VII. FINAL REMARKS

If I am right in claiming that the biconditional doctrine and the truth doctrine are dogmas, then the ultimate theorems of a theory of meaning do not need to take the extensional form '*s* is *T* if and only if *p*' and the use of the predicate 'is *T*' is merely an expedient in the recursive machinery of the theory. Pursuers of Davidson's program can coherently claim the notion of truth to have no explanatory significance in semantics. Where does that leave plausible-sounding slogans such as

(S1) The meaning of a sentence is its truth condition

and

(S2) To know the meaning of a sentence is to know under what conditions it would be true.

Is the label "truth-conditional semantics" a misnomer?

The answer is that the terminology of truth conditions is indeed partly misleading. (S1) is informative only on the background of Davidson's biconditional dogma. It is correct insofar as it expresses the underlying Davidsonian insight that one can construct a theory of meaning for a language which explains the compositionality of meaning by exploiting recursive techniques first devised by Tarski to define truth in a formal language. But often (S1) is misleadingly used to suggest a deeper significance. (S2) is true only insofar as it expresses the insight that to know the meaning of a sentence is to know how to use it correctly. In some typical cases, it is correct to use a sentence only if its content is true, so knowing under what conditions it would be true (whatever that means) will often help one to use it correctly. But this insight is quite independent of Davidson's main idea in "Truth and Meaning," namely, to use a theory of a certain recursive structure as a theory of content for a language. It rather belongs to a theory of communicative (illocutionary) acts, which might state, for example, that one should assert that *p* only if *p* (= that one should assert that *p* only if it is true that *p*).[28]

But is it not still too much of a coincidence that it should be possible correctly to use the predicate 'is true' in the intermediate lemmas of the theory, if in fact any uninterpreted predicate would have served just as well?

---

[28] See the discussion in the previous section for more accurate thoughts on assertion.

The answer is that this is not a coincidence. It can be neatly explained by minimal (deflationist) assumptions about the function of the truth predicate without uncovering any deep link between truth and meaning.[29] Deflationists about truth claim that truth is not a mysterious property. All we need to know about truth is encapsulated in the way the truth predicate solves a simple syntactic problem. The problem arises, for example, when there are sentences (or propositions) that we can mention but not use (which we cannot make explicit). For example, many people are unable to state Fermat's theorem. Nevertheless, they can easily make reference to it by just using the expression 'Fermat's theorem'. (They often know that Fermat's theorem has only recently been proven.) Merely making reference to the theorem, however, does not yet allow one, for example, to assert the theorem or to use it as the antecedent of a conditional. This is where the truth predicate is useful. It is governed by the rule that by applying it to the name of any sentence (or proposition), one gets a new sentence which is equivalent to the sentence of which truth was predicated (or which expresses a proposition equivalent to the proposition of which truth was predicated). Some have expressed this by calling the truth predicate a prosentence forming operator.[30] For example, the sentence 'Fermat's theorem is true' expresses a proposition equivalent to Fermat's theorem. Another typical example is the sentence 'All of Davidson's doctrines are true', which is equivalent to the conjunction of Davidson's doctrines but much easier to state.

This simple function of the truth predicate explains why it should be possible to interpret the predicate used in T-theorems as the truth predicate. In generating our meaning-specifying theorems from semantic axioms, we need to pair structural descriptions of object-language sentences with their metalanguage translations. In temporarily "filling the gap," we ought to use the material biconditional, as that helps our derivations (via a rule of substitution of equivalents, see section III above). Since a structural description of a sentence cannot flank the biconditional, we need to complete it with some predicate 'is *T*' to form a sentence. Since the other side of the biconditional is that sentence's translation (if the theory is correct), it is no coincidence that interpreting 'is *T*' as the truth predicate makes the T-theorems come out true For as we have seen, the truth predicate forms sentences that are equivalent to the sentence of which it is predicated The usefulness of Tarski's methods in a compositional theory of meaning does not, therefore, indicate any deep connection between the notion of truth and that of meaning

---

[29] As done by Michael Williams in his "Meaning and Deflationary Truth," this JOURNAL, XCVI, 11 (November 1999): 545-64, p. 557.

[30] For example, Brandom, in *Making It Explicit*, chapter 5.

MAX KÖLBEL

New Hall/University of Cambridge

SELF-DEFEATING SYMBOLISM IN POLITICS*

Many people seem insensitive to the fact that the public policies they support might frustrate the values in the name of which they support those policies For example, many people support minimum-wage laws by invoking the plight of the poor while simultaneously believing that those laws hurt the poor Some writers have argued that only a nonstandard, symbolic model of rationality can vindicate this puzzling behavior as rational On their view, many people support minimum-wage laws to symbolize or express concern for the poor For these authors, political theory has overlooked the fact that many people act, not to achieve some identifiable political end, but to express a commitment [1] They do not act instrumentally but expressively or symbolically In our example, the agent who supports minimum-wage laws in the name of helping the poor yet believes that those laws hurt the poor does not aim primarily at helping the poor but at expressing symbolically her concern for the poor On this view, she is rational despite the fact that she seems to act counterproductively

Our aim is to show that ordinary models of rationality provide the best account of self-defeating political behavior Many people whose political behavior would seem incomprehensible given their stated aims and values are actually displaying an ordinary, instrumental sort of rationality We start by offering characterizations of symbolic behavior and self-defeating behavior in politics (section I) We then discuss Robert Nozick's account of political symbolism and of symbolic and causal utility (section II) We argue in section III that most symbolic acts are instrumentally rational, and we distinguish in section IV among various types of political self-defeatingness and disclose their instrumental rationality We use ordinary rational-choice as-

---

* We thank several audiences who heard earlier versions of this paper the Arizona State University Moral, Political, and Legal Philosophy Discussion Group, the Department of Philosophy at the University of Arizona, participants in the Law and Normativity Conference at the Universidad Torcuato Di Tella, June 1999, and participants in the Colloquium on Practical Philosophy, Argentine Society for Philosophical Analysis, June 2001 We owe special thanks to Brad Armendt, Geoffrey Brennan, David Copp, Thomas Pogge, Joseph Raz, Eduardo Rivera López, and Seana Shiffrin

[1] See Robert Nozick, *The Nature of Rationality* (Cambridge Harvard, 1993), pp 25-30, Elizabeth S Anderson and Richard H Pildes, "Expressive Theories of Law A General Restatement," *University of Pennsylvania Law Review*, CXLVIII (2000) 1503-75, and Cass Sunstein, "On the Expressive Function of Law," *University of Pennsylvania Law Review*, CXLIV (1996) 2021-53

0022-362X/01/9812/636–52

sumptions to show in section v that ignorance, error, and insincerity explain most self-defeating political behavior We set forth our conclusions in section vi

I SYMBOLIC BEHAVIOR IN POLITICS

We define *symbolic behavior* as behavior intended by the agent to express adherence to a principle without such behavior falling under the principle[2] We are interested here in the relationship between symbolic behavior and some kind of principle, especially of a moral sort An agent may instantiate the principle or may express his support for it or do both For example, suppose that I consider myself bound by the principle, 'Be faithful to your spouse' I instantiate this principle when I refuse to yield to temptation of betrayal (it is unclear whether I instantiate the principle every time I fail to betray) In contrast, I express support for the principle when I say how important fidelity is, or when I give my spouse a wedding anniversary gift Perhaps one can say, with Nozick, that whenever I instantiate a principle I also express support for that principle (*op cit*, p 33) Be that as it may, the relationship does not seem to work the other way around the fact that I have expressed support for a principle does not mean that I have instantiated the principle My symbolic or expressive[3] behavior is not yet compliance with duty The symbolic act does not cause the state of affairs recommended by the principle Rather, it expresses approval of, or appreciation for, the principle itself

Symbolic meaning may be *social* or *subjective* In the first case, the agent is doing something that is conventionally regarded as showing adherence to a principle In the second case, the agent is doing something that shows to himself that he is adhering to a principle The agent who subjectively expresses adherence to a principle is endorsing the principle The agent whose behavior is socially seen as symbolizing a principle may or may not endorse the principle

We confine our discussion to symbolic behavior in politics, more specifically, to (seemingly) self-defeating attitudes toward laws, policies, and other public measures Public measures have social consequences Yet symbolic behavior is not about recommending measures because of their valuable social consequences but about expressing support or adherence to a principle So an agent who supports a law for symbolic reasons is expressing adherence to the (moral) principle that presumably informs that law She may or may not will the causal consequences that ensue from that law A typical political agent *A* who acts symbolically

---

[2] Similar definitions result from referring to values or goals rather than to principles

[3] We use 'symbolic' and 'expressive' interchangeably, although the distinction between the two might be useful in a different context Thus, we may say that someone may want to express herself without communicating anything—this may be called *expressive behavior* *Symbolic behavior*, on the other hand, can be defined as communicative The difference matters for a distinction we make below between self-defeating reformers and self-defeating *performers*—that is, persons who seem to act counterproductively but want only to express themselves

(1) publicly expresses adherence to principle *P* by supporting law *L*

We shall say that *A* supports *L* in the name of *P* A self-defeating symbolic political agent is one who recommends a public measure (such as a law) in the name of a principle that, so he believes, would be frustrated (either directly or in virtue of behavior performed in response to the threat of sanctions) by the enactment of that law A self-defeating political agent, therefore, in addition to (1),

(2) believes that *L*, if enacted, would frustrate the realization of *P* (that is, would frustrate the state of affairs enjoined by *P*)

Call this agent a *self-defeating reformer* (SDR) Note that the sense in which this agent's behavior is self-defeating is that she believes that the law she supports would frustrate the realization of the principle in the name of which she acts [4] She does not necessarily believe that *her act of support* is detrimental to the principle Nor do we assume in this definition that she endorses *P* All we assume is that her behavior may conventionally express adherence to *P*, as we noted earlier, symbolic meaning may be social, subjective, or both Someone may object that agents who satisfy (1) and (2) need not be really self-defeating For example, an insincere agent (that is, one who does not care about *P*) would not be really self-defeating This objection, however, leaves untouched the puzzle that we want to examine, namely, why rational persons would support counterproductive political measures, whether or not we call them self-defeating agents [5]

II SYMBOLIC AND CAUSAL UTILITY NOZICK'S CHALLENGE

In *The Nature of Rationality*, Nozick claims that people who perform an act to express adherence to a value while simultaneously believing that the act frustrates the realization of the value can nonetheless be acting rationally (*op cit*, pp 26-28) They can be rational because they derive symbolic utility from that behavior Nozick thinks that in working exclusively with causally derived utilities (and thus ignoring symbolic utilities), standard decision theory is incapable of uncovering the rationality of such people

---

[4] For stylistic reasons, we shall refer to agents who believe such propositions, meaning to include those who believe that such propositions are plausible Our argument holds for both cases

[5] Indeed, as we shall argue, most SDRs are instrumentally rational, so they are not self-defeating in any sense that denies instrumental rationality The self-defeating quality of their behavior is that the political measures they support would, if enacted, frustrate the values they invoke to justify such support

Nozick uses two examples from politics We already mentioned the first someone supports minimum-wage laws in the name of the principle that we ought to help the poor, even when he believes that those laws actually hurt the poor[6] He still does so because the act of supporting minimum-wage laws symbolizes concern for the poor He derives utility from expressing support for that principle ('We ought to help the poor'), quite apart from whether or not those laws help the poor[7] In the second example, someone supports anti-drug measures without paying attention to evidence that those measures increase drug consumption and crime Here again, the act symbolizes the moral condemnation of drug use and the crime associated with such use (What could be more symbolic of the revulsion for drug use than prohibiting it?) In both cases, the agent derives symbolic utility from supporting those measures, notwithstanding his belief that they may prevent the (subjectively) valuable outcome that his support is symbolizing

Nozick claims that symbolic rationality cannot be reduced to instrumental rationality He endorses the standard view that rational agents maximize utility or value, but distinguishes between *causal* and

---

[6] Sunstein also uses that example, though in a normative context He concludes that people who believe that minimum-wage laws hurt the poor should not, as a matter of political morality, support them just for symbolic reasons (*op cit*, pp 2046-47) In their defense of expressive theories of law, Anderson and Pildes nonetheless say that the view that expressive evaluation somehow requires people to ignore the consequences of action is absurd (*op cit*, p 1513)

[7] We use the example of self-defeating support of minimum-wage laws, where the agent whose rationality we assess happens to adopt a view against free-labor markets The argument here does not depend on the proposition that minimum-wage laws hurt the poor or, for that matter, on any pro-market or libertarian political view For a brief survey of the pros and cons of minimum-wage laws as far as low-wage workers are concerned, see Campbell R McConnell and Stanley L Brue, *Economics* (Boston McGraw-Hill, 1999), pp 594-95 What is important here is that the self-defeating agent *believes, or at least regards as plausible,* that minimum-wage laws hurt the poor One can easily think of examples of pro-market or libertarian self-defeating agents Imagine someone who opposes deregulation of utilities in the name of efficiency or wealth creation Suppose further that economists convince him that, because of market failure, the deregulation he proposes will decrease general wealth, that is, it will produce inefficient outcomes Yet this agent may persist in his support of deregulation if he believes that such behavior socially symbolizes commitment to efficiency In our terminology, this person is a self-defeating political agent

symbolic utility[8] Symbolic utility is a further strand, a distinct, noncausal, concept The agent performs an act because it has symbolic value for him and others That the act is symbolic means that it stands for something else—a principle, or value, or something desirable Crucially, however, the act does not cause the state of affairs recommended by the principle, the value, or the desirable thing What the act symbolizes (that "something else") has value for the agent, and that value is imputed back to the act, thus "giving *it* greater utility than it appeared to have" (*op cit*, p 26) Nozick recognizes, of course, that utility can be imputed back along causal lines action *A* may have value for the agent because she believes that *A* causally contributes to an outcome *O* that she (instrumentally or ultimately) values The metaphor of "imputing back" attempts to show that the value of *A* is parasitic on the value of *O* For Nozick, then, *A* is valuable either because it causes *O* or because it symbolizes *O* In the latter case, *A* borrows value from *O* because *A* stands for *O* This will be true even when *A* does not contribute causally to *O*—in fact, it will be true, according to Nozick, even when *A* frustrates the realization of *O*

Nozick observes that standard decision theory cannot easily explain symbolic utility and thus cannot explain what he views as the anomalous case the SDR case An SDR believes that the policy he advocates frustrates the realization of the symbolized principle Nozick claims that an SDR acts irrationally under standard decision theory, yet he argues that we can salvage his rationality by appealing to irreducibly symbolic utility For Nozick, traditional decision theory has overlooked the fact that symbolic utility is independent of causal utility an act's utility may be noncausal—purely symbolic, as it were Its value for the agent does not depend, even partly, on the probabilities of its causing future situations that actualize the symbolized value Nozick thinks his view is supported by the fact that many people persist in doing something "in the face of strong evidence that it does not actually have the presumed causal consequences" (*op cit*, p 27) The symbolic, rather than the (putative) causal connection to an outcome plays a central role in the decision to do it In our example, an SDR derives symbolic utility from supporting minimum-wage laws in the name of the principle that one ought to help the poor, even though he believes that those laws will hurt the poor Nozick observes that perhaps a rational agent should choose a different act to express his adherence to the principle, one that does not have these bad consequences But simply refraining from performing that act (that is, without replacing it by another symbolically appropriate act) would deprive people of its symbolic utility—something, Nozick thinks, they are unwilling to let happen (*op cit*, p 27)

---

[8] Nozick discusses a third kind of utility as well—evidential utility (*op cit*, pp 45-59) Our concerns in this article require discussion of symbolic and causal utility alone

Why does Nozick believe that his account improves on traditional decision theory? He may be assuming a plausible adequacy condition for a theory of rationality the theory must depict most behavior as rational [9] Presumably, a theory of rationality under which most people were crazy would not be attractive So Nozick feels compelled to introduce a symbolic, nonstandard form of rationality in order to account for the inordinate number of persons who seem to behave in a self-defeating way in the political arena We believe, however, that such behavior can be explained by resorting to standard, instrumental accounts of rationality

III SYMBOLIZING AS THE INTENDED OUTCOME

We interpret judgments about rationality as relying on testable hypotheses about an agent's means and goals We also assume that people who defend symbolic behavior use testable hypotheses as well We thus avoid the use of ad hoc or circular notions of utility That is, we shall not say that someone who prefers to do x rather than y is rational because he *eo ipso* derives more utility from x than from y In other words, in order to assess people's rationality we do not simply work with their revealed preferences we have to know more about their goals and beliefs We avoid ad hoc judgments of rationality

We do not believe that in order to explain self-defeating political behavior we should amend standard decision theory, which evaluates and explains behavior within the framework of actions causing outcomes A rational agent believes that her action will (be likely to) cause an outcome that she values This view of rationality is, therefore, instrumental a rational agent chooses what she regards as efficient means to her goals But surely there is a sense of "action" in which *the action itself* can be the desired outcome An agent may derive utility or value from doing A, where *his doing it* is the outcome he intends He may take an interest in his own action, as something different from further causal consequences of the action It is natural for him to believe that doing that action is an efficient means to satisfy such interest For example, many people value playing tennis quite independently of further causal consequences that they expect to achieve (for example, improved health, better bodily appearance) So the fact that an action is the very outcome sought by the agent does not prevent standard rational-choice theory from explaining the agent's behavior the action itself has utility or value for the agent The agent who acts symbolically, then, draws value or utility from expressing his adherence to a principle His expressive act is what he values

---

[9] We use the terms 'decision theory' and 'theory of rationality' interchangeably

Nozick observes that often an action is valuable (has utility for) the agent because it symbolizes something valuable, and so the action borrows utility from the symbolized thing The value of the symbolized thing is imputed back to the action We believe, however, that this does not undermine the basic analysis of rational-choice theory the agent draws utility from the state of affairs he causes Sometimes the causal chain is very short the agent who derives symbolic utility from a ritual dance causes the relevant outcome (that is, his dancing) in a very straightforward way But this should not be more mysterious than deriving utility from more distant outcomes, as in dancing for money The two examples differ only in the lengths of the respective causal chains, and this difference is irrelevant to decision theory What complicates the issue in Nozick's analysis is that there are two kinds of valuable outcomes in his examples One is the action itself (from which the agent derives value directly, as it were), and the other is the principle symbolized by the action Presumably, the agent would say that the reason why he values the action is because it symbolizes that principle (for example, the duty to help the poor), but the fact remains that the agent derives utility from his doing the symbolic action Nozick's challenge to traditional decision theory trades on the fact that there is no causal relation between the action and the instantiation of the principle in the name of which the action is undertaken The suggestion is that we need to postulate a new relationship between action and outcome to explain what is otherwise mysterious We think that this view overlooks the fact that there is a causal relation, a relation of means to ends, and thus no mystery, once we acknowledge that the symbolizing is itself the end

Another way to put this is as follows We value things and events because of some of their properties Sometimes such properties are causal in character, as when we value a political revolution because it brought about justice But this need not be so I may value a car because it *is* fast Moreover, we often value something because of its noncausal relational properties, for example, I may value John because he is a member of the Red Cross Shall we say that I impute back to John the value I ascribe to the Red Cross along membership lines? Yes, provided that this does not mean anything different from saying that I impute back to the car the value that I find in the car's speed (along a nonrelational property) In both cases, a property grounds a value judgment and renders rational any cost-efficient undertaking aimed at instantiating that property (for example, my producing such a car, my advising John to become a member of the Red Cross). Similarly, the fact that I value *S* because it symbolizes *T* renders it rational, in the standard sense, for me (cost efficiently) to bring about *S*. What matters here is that sometimes my own actions instantiate *S*. As David Christensen[10] says, "the utility comes from the *symbolizing* of the situation," from the act of communicating something to others (*ibid.*, p. 265). Nozick is correct that this utility is independent of the expectation of the symbolized situation coming to pass, but his assertion that the utility flows from the situation back to the act wrongly suggests an anomaly for standard decision theory. It is not that the symbolized situation generates utility for the agent but rather that the symbolic act does. When someone acts symbolically, we do not know yet if he is really prepared to honor the principle on behalf of which he acts. All we know is that he values his (social or subjective) expression of support for the principle.

As we anticipated, we assume that judgments about rationality are not ad hoc. So the foregoing account of symbolic behavior requires us to have independent evidence about the agent's utility functions. In the dancing example, attributing to the agent the ultimate end of dancing would be ad hoc if we did not have independent evidence that he has such end. But, as we said, we are not defending the rational-choice model in this way. In order to say in an interesting way that the dancing itself is the agent's ultimate end, we need to have independent evidence: he might have paid for dancing lessons in the past, he might have told his friends that he loved dancing, and so on. Since conceivably the available evidence may tell against such a hypothesis, our analysis makes room for irrational behavior. As we shall argue in section v, however, most self-defeating political behavior can be analyzed in terms of testable rational-choice assumptions.

IV. IS THE SDR RATIONAL?

The ordinary rationality of SDRs can be best seen by distinguishing several cases of apparently self-defeating political behavior. Imagine an agent who sincerely endorses the principle that one must help the poor. He supports minimum-wage laws, yet believes that these laws, considered in isolation, would hurt the poor. He does not believe, however, that *his act* will hurt the poor, all things considered. He chooses publicly to support minimum-wage laws because he thinks that he thereby instantiates a symbol that, in turn, starts a causal chain resulting in the betterment of the poor. He may or may not derive utility from symbolizing the principle he endorses. But he does believe that his act (that is, his public support of those laws) will have other consequences that are beneficial to the poor: people will be encouraged to give more to charity, or will support public causes or politicians who are committed to helping the poor. Characteristically, this agent believes that the measures he supports are counterproductive in terms of the value he cherishes, but that *his support* for those measures is not. The agent we are imagining adds the expected symbolic utility (if any) of his supporting the laws to the expected causal utility of those other consequences, and subtracts the expected utility of the possible enactment of the laws he supports. (The latter expected utility is negligibly negative, given that public support of laws is normally causally irrelevant to their being enacted.) If the balance is positive he performs the symbolic action.[11] One major reason why a rational agent will be tempted to perform the symbolic act is that it is within his power, something he can do for sure, here and now (as opposed to bringing about the improvement of the poor, which is outside his power).

---

[10] "Book Review," *Noûs*, xxix (1995): 259-74. See also Gerald F. Gaus, "Goals, Symbols and Principles: An Evaluation of Nozick's Theory of Practical Rationality," presented at the Colloquium on Political Philosophy held at Universidad Torcuato Di Tella Law School, Buenos Aires (June 7, 2000), in David Schmidtz, ed., *Nozick* (New York: Cambridge, forthcoming).

So this agent, in addition to satisfying the dyad (1)-(2) above, meets the following two conditions:

(3) He endorses *P* as his overriding principle.
(4) He thinks that his supporting *L* (as something different from *L* itself) will advance *P*, all things considered.

Call this person the *consistent agent*. It seems safe to conclude that this agent is rational because he does not think that his behavior, as something different from the laws themselves, harms the poor, all things considered. He maximizes his utility by doing what he believes is causally conducive to that end, so he is instrumentally rational.

But suppose an agent is not primarily interested in the causal consequences of his symbolizing help to the poor. Conceivably, his overriding goal may be the symbolizing itself. Normally, this agent (the *expressive agent*) will maximize his utility quite straightforwardly—he just does the symbolic act (for example, he wears the ribbon, he publicly supports a principle, and so on). We can say that the utility derived by the expressive agent is symbolic in the sense that he performs an action because of its being a symbol of something (socially or subjectively) valuable. Under our definition of symbolic behavior, the agent who acts symbolically does something in the name of, or for the sake of, a value or principle without instantiating that principle. The agent who acts symbolically brings about something that is valuable to him, namely, an act that has symbolic meaning (that is, that stands for something else, however we construe such relation). Accordingly, symbolic behavior possesses the same instrumental structure that we find in nonsymbolic contexts. The agent causes the symbol (the act) to occur (brings it about, performs it), just as an agent who derives utility from dancing causes (brings about, performs) acts falling under the description 'his dancing'.

---

[11] We employ the terms 'addition', 'subtraction', and (later) 'algebraic sum' loosely, for terminological convenience. We do not mean to endorse a cardinal view of utility. But our account does assume intrapersonal ordinal rankings of utility. 'Addition' and 'subtraction' allude to the agent's subjective balancing of alternatives.

It might be useful to compare the consistent agent with the expressive agent. Both the consistent agent and the expressive agent are SDRs: they satisfy the above propositions (1) and (2). The consistent agent is attentive to consequences, since he values the fact that his symbolic behavior, apart from having utility or value in itself to him (and maybe others) does (in his view) help the poor, all things considered. By contrast, the expressive agent is less attentive to those consequences: he does not want, primarily at least, to help the poor. Unlike the consistent agent, the expressive agent does not believe that his behavior helps the poor, all things considered. He satisfies neither (3) nor (4) above. He derives utility from supporting minimum-wage laws because his support symbolizes the duty to help the poor. He derives utility from an instance of this symbolic connection (from being on the right side, as it were). We may say that the utility he derives from the act of supporting minimum-wage laws exceeds whatever (negative) utility he derives from the enactment of minimum-wage laws, discounted by the probability of such enactment given his support. Of course, under normal circumstances such expected (negative) utility is very low because his support is causally irrelevant. So in ordinary cases the cost to the expressive agent of being on the right side (measured by the expected disutility, if any, he might derive from the fact that the poor will be hurt as a result of his present behavior) is very low. The consistent agent, on the other hand, sincerely cares about the poor. Given his belief that the laws are unlikely[12] to be enacted, he may safely support minimum-wage laws

---

[12] Politicians, and especially legislators, may significantly influence the outcome. But for very rare cases of irrationality, standard decision theory can explain their behavior along the lines we shall suggest in section v.

The upshot of all this is that an SDR remains rational in the sense of traditional decision theory Neither in the consistent-agent nor in the expressive-agent cases, need we assume that utility is noncausally imputed back to current behavior In particular, there is no "imputing back" along symbolic connections The consistent agent and the expressive agent are, therefore, rational under standard decision theory They only differ in the value they assign to the symbolic action vis-à-vis its consequences

To be sure, these agents' behavior is puzzling because the policies they support are counterproductive in terms of the values they ostensibly invoke Here again we should distinguish between the consistent agent and the expressive agent The consistent agent's behavior coheres he believes he ought to help the poor and he knowingly behaves to achieve just that By contrast, the expressive agent's behavior may be incoherent If he actually endorses[13] the principle in the name of which he acts, his behavior is incoherent he believes he ought to help the poor but he knows that his behavior will not achieve that end To the extent that he values helping the poor more than he values the symbolizing itself, he will be instrumentally irrational But if the expressive agent does not endorse as paramount the principle in the name of which he acts, his behavior coheres he does not believe that what he does will help the poor, but helping the poor is not his primary goal His primary goal is the symbolizing itself

What about people who (a) advocate policies that they foresee will be detrimental to their values, (b) have as their reason for such advocacy the fact that those policies (or advocacy of them) express a commitment to those values, but (c) do not behave in that way *in the name of realizing* such values? Consider this case A woman stands by her sick son's bed She knows that his chances of recovery are small Terribly moved by her son's plight, she embraces him, knowing that this will kill him immediately (the *poignant case*[14]) What are we to make of the mother's rationality? She is not an SDR She does not fit our earlier definition of self-defeating behavior, because she is not supporting a course of action in the name of things she deems valuable Perhaps we can attribute this mother the intention to express adherence to a value, such as the value of having healthy children, and in this sense hers is a symbolic action But, unlike the consistent agent and the expressive agent, what she does is something other than supporting specific measures in the name of the value

---

[13] Recall that we did not stipulate that an SDR necessarily endorses the principle in the name of which she acts See proposition (1) in section I

[14] We are grateful to Geoffrey Brennan for suggesting this example

Call this behavior *symbolic** to distinguish it from the sort of symbolism discussed so far and defined at the outset of section I Unlike the act of supporting something, the mother's embracing her son is symbolic* in that it lacks propositional implications, so it makes no sense to assess it in terms of how favorable or detrimental to (nonexpressive) goals her act is We can call this mother a *self-defeating performer* (SDP) Because an SDP is, as such, not supporting or advocating anything, she is not an SDR

The mother in the poignant case satisfies the conditions laid down for *X* in the following dyad

(I) *X* believes that her doing *A* symbolizes* valuable thing *T*
(II) *X* believes that *A* causes outcomes incompatible with *T*

Note that, unlike the formulation offered in section I, this one makes no reference to *X*'s supporting measures that frustrate her ends So her behavior should not be confused with that of the expressive agent While the expressive agent primarily values his symbolizing, his behavior has a causal component in that he supports political measures, and this support is conventionally understood as entailing that those measures are instrumental to the realization of the principle in the name of which he acts Thus, when confronted with someone who supports minimum-wage laws in the name of helping the poor, most people will reasonably infer that this agent believes that minimum-wage laws help the poor On the other hand, an SDP, like the mother, displays purely expressive behavior, as it were, so there is no potential for incoherence In short, her expressive behavior is symbolic*, not symbolic Unlike an SDP, an SDR supports a policy in a political context As such, he takes a moral stance in the sense that he is committed to justifying his views if need be This makes room for incoherence—the SDR acts in the name of principles whose realization may be frustrated by the policy he advocates To be sure, the SDP is still acting in a kind of self-defeating way because she knowingly engages in behavior that frustrates her values But she does not instantiate the symbol in the name of a principle, she just expresses her values, as it were

Is the SDP instrumentally rational? There is no definite answer all depends on how we fill out the description of the case If, in the poignant case, the mother's overriding end is to help heal her son, then she is arguably irrational in the standard sense she fails to maximize her utility given her beliefs[15] If the expected utility she derives from embracing her son is greater, then she is rational in the same sense

---

[15] Remember that we are using a noncircular notion of "utility" (see section III)

Can ostensibly self-defeating political behavior be modeled in terms of the SDP? If the answer is yes, then prima facie counterproductive political agents would not be displaying self-defeating behavior, and a fortiori they would be instrumentally rational As we have just seen, unlike the SDR, the SDP is not supporting anything She expresses adherence to a principle or value, but, to the extent that she does not advocate measures in the name of a principle or value, there is nothing in her behavior to be inconsistent with If this were the correct model for political agents (for example, the supporter of minimum-wage laws and drug prohibition), then they would not be self-defeating, and their symbolic* behavior would not allow us to decide whether they are rational or not Contrary to what seems a natural interpretation of their conduct, those people would not be properly supporting anything it would be more accurate to view them as *embracing* a valuable cause [16]

We doubt, however, that many agents whose political discourse sounds counterproductive are really SDPs Embracing a cause is not like embracing a person Support of political measures is not simply expressive of commitment to values or principles, it is also performed in the name of values or principles As we said, an agent who supports laws in the name of values or principles is committed to endorsing causal beliefs Thus, support of minimum-wage laws in the name of the principle that we ought to help the poor conventionally implies a commitment to the causal belief that such laws help the poor (or, at the very least, that those laws do not harm the poor) Contrast this political attitude with that of the mother in the poignant case She may be expressing her love for her son, but she is not behaving in the name of a principle that mandates helping her son survive So we conclude that the only irrational political agents are those SDRs who sincerely endorse as of paramount value the principle in the name of which they act We suspect that these agents are very rare [17] Let us substantiate this suspicion in the next section

V WHY POLITICAL AGENTS APPEAR TO NEGLECT CONSEQUENCES

As we pointed out in section II, we expect a theory of rationality to explain the way most people behave But we also required of such explanations not to be ad hoc Therefore, we need to offer independent evidence of why in politics a surprisingly large number of people seem to act symbolically, even at the cost of self-defeatingness. We showed that symbolic political agents typically value their act of symbolizing, but not necessarily the symbolized thing. They need not care much about the consequences of the laws they support for symbolic reasons. This will be true whether or not their behavior is self-defeating, that is, whether or not they believe that the laws they support would, if enacted, frustrate the principle in the name of which they act. What could then possibly motivate people to act symbolically in politics in disregard for the social consequences of the policies they advocate?

---

[16] Symbolism* may play a role in moral assessment Depending on the circumstances, symbolic* reasons might exonerate otherwise objectionable behavior like the mother, political agents would be moved by terrible facts (for example, unsatisfied dire needs, violence)

[17] So is the poignant case, and for this reason it should not threaten standard decision theory, even assuming that the mother is typically irrational

We offer two conjectures. The first is that in many cases people whose behavior may appear self-defeating are ignorant of, or mistaken about, the relevant causal connections. Suppose someone supports minimum-wage laws believing that they will help the poor. This agent is not self-defeating: he does not satisfy (2) above, that is, he does not believe that the measures he supports will frustrate the goals in the name of which he acts. This person, however, might seem to act in a self-defeating way because, on a plausible reading of economic theory, minimum-wage laws might well hurt the poor. His apparent self-defeatingness is even more noticeable when he is unwilling to consider relevant evidence and arguments. Obviously, if many people neglect relevant evidence, chances are that false beliefs about society will be pervasive, and this in turn may lead those people to advance seemingly counterproductive political proposals.

Why would so many people hold mistaken social theories? The answer, we suspect, is that political interaction creates incentives for people to err. Consider how political scientists explain the fact that voters do not invest much in political information. Given the negligible effect of a vote in an election (a vote is consequential only when it breaks a tie) and the complexity of the issues involved, a voter's expected utility of investing in information about such issues is also negligible, if not negative. In short, voters tend to remain ignorant about politics, and rationally so.[18] Instrumentally rational voters will fail to invest in political information. In our example of minimum-wage laws, even a voter who is morally motivated to help the poor will remain ignorant of relevant facts and theories about effective political ways of helping the poor. It certainly makes more sense for him to help the poor through charity, for example, than trying to help the poor through his informed participation in the political process Voters will predictably rely on quickly and easily understandable theories and "facts," because the personal cost of grasping them is low They will normally rely on common-sense theories and cheap information (for example, television shows) in deciding whom to vote for In addition, they are more likely to believe in theories that postulate short, simple, or vivid causal chains (like the thesis that minimum-wage laws force the employer to raise salaries and thus improve the lot of the poor) than theories that are harder to grasp These include counterintuitive theories and those which postulate long or complex causal chains that involve positive sum games and dynamic effects of social interaction The economist's claim that minimum-wage laws hurt the poor relies on such complex theories It seems safe to say that here cost parallels difficulty learning easy theories is cheaper than learning complex theories [19] Yet there need not be correlation between simplicity and truth

---

[18] See the classical work on rational ignorance by Anthony Downs, *An Economic Theory of Democracy* (New York: Harper, 1957), especially part III, pp. 207-76.

Now consider politicians They have an electoral incentive to know these facts about people's beliefs Hence, they will predictably employ rhetoric based on short or simple causal chains—otherwise, they will find themselves competitively disadvantaged in the struggle for rationally ignorant votes Indeed, rational error stems from the same logic as rational ignorance The rational behavior of both politicians and voters explains why certain general views about how society works tend to prevail Differential incentives to acquire and spread political information lead citizens to accept or reject social theories through a truth-insensitive process Under these conditions, and assuming that sometimes the right explanation of a social phenomenon is hard to understand, we should expect widespread acceptance of false theories The upshot is that we do not have to appeal to symbolic reasons to explain the apparently self-defeating public positions held by many citizens Neither do symbolic reasons account for the politicians' truth-insensitive[20] rhetoric they are utility maximizers in a causal sense whatever their utility functions (for example, maximization of political power, or votes[21]), and often choose to lie in the pursuit of their goals[22] So, prima facie SDRs are likely to be rationally ignorant or mistaken if they are voters, whereas they are likely to be rationally insincere if they are (successful) politicians We reach this conclusion without assuming that utility can be imputed back along symbolic lines We rely instead on standard rational-choice assumptions

---

[19] Difficulty and cost need not go together See G A Cohen, "If You're an Egalitarian, How Come You're So Rich?" in Guido Pincione and Horacio Spector, eds , *Rights, Equality, and Liberty* (Boston Kluwer, 2000), pp 23-26

[20] Politicians do have an incentive to care about truth when falsehood is transparent to voters Our point is that political competition provides politicians with an incentive to care about truth in this incidental sense only

[21] Such utility functions are compatible with ascribing public-spirited *ultimate* ends to the politicians We do not take sides here with any particular view about the contents of those ends

The second conjecture is that many people have incentives to appear on the side of what is socially perceived as right Being so perceived is usually associated with various kinds of benefits success at work, acceptance in certain circles (political or otherwise), and in general the psychological and material rewards accruing from social prestige An SDR who wants to appear on the right side might or might not be concerned about his self-defeatingness That is, he might or might not care about the fact that his support, say, for minimum-wage laws in the name of helping the poor is coupled with his belief that those laws hurt the poor We can imagine an SDR reasoning as follows "I know that minimum-wage laws will hurt the poor, but I support them because my overriding goal is to be seen by others as promoting those laws out of concern for the poor So I pretend that I do not know that those laws hurt the poor (otherwise, my support would not have such symbolic meaning) I would prefer that the poor not be hurt (and probably they will not be hurt because the likelihood of enactment is low), but at any rate I draw value from the social approval I get thanks to the symbolic significance that I convey by my action " Let us call this agent the *posturer* He seeks social esteem as his primary goal, and uses his expressive behavior to earn it He values his expressive behavior instrumentally, and in this he differs from the expressive agent (see section IV) To the extent that the posturer does not believe that he has poor-improvement grounds for supporting laws, the coherence of his behavior is saved Insincerity sustains coherence He asserts, but does not believe, the following proposition, entailed by his advocacy of minimum-wage laws in the name of the poor 'Minimum-wage laws help the poor' He also asserts (or implies) that his goal is to help the poor, but he does not really have that goal, at least as his overriding goal This agent is thus doubly insincere

---

[22] Rational-choice models of voting behavior have difficulty explaining voter turnout Why do so many people vote, given that, for reasons similar to those underlying the rational ignorance effect, the expected utility of voting is negligible (or even negative, given the cost of going to the polling booth)? Brennan and Loren Lomasky introduce expressive reasons to model this behavior In a modern democracy, because one vote cannot make a difference, symbolic reasons may play a role in explaining voter turnout—see Brennan and Lomasky, *Democracy and Decision* (New York Cambridge, 1993) For an unsympathetic survey of the literature on rational-choice explanations of voter turnout, see Donald P Green and Ian Shapiro, *Pathologies of Rational Choice Theory* (New Haven Yale, 1994), pp 47-71 Notice, however, that expressive models of voter turnout are compatible with the analysis of rational ignorance and error offered in the text Perhaps symbolic utility draws the voter to the polling booth, but her vote will be still rationally ignorant or mistaken Whatever the best explanation of voter turnout may be, we believe that the rational ignorance/error effect accounts for the behavior of sincere and rational individuals who look like SDRs We need not appeal to symbolic reasons

Does the posturer's insincerity affect his rationality? We do not think so Agents who are insincere in this way are instrumentally rational The instrumental rationality of an act does not turn on the means employed by the agent to bring about his preferred outcome Insincerity, in particular, may be the agent's means to convey the symbolic message associated with his support of minimum-wage laws Only if sincerity is a definitional feature of the symbolic act can insincerity frustrate symbolic meaning But there is no substantive constraint (including mental constraints, like sincerity) in successful symbolizing [23] As we saw, an agent may instantiate a symbol without endorsing the symbolized value or principle So there is nothing in rational behavior that rules out insincerity We conclude, then, that ostensibly counterproductive behavior whereby the agent gets social esteem by being insincere is instrumentally rational [24]

VI CONCLUSION

Most people who engage in self-defeating political symbolism display an instrumental sort of rationality Symbolic agents typically value instrumentally the act of symbolizing When seen in this light, their behavior is no longer puzzling To be sure, many people seem simultaneously to attach overriding force to the principle they invoke while behaving counterproductively with respect to that principle If this appearance turned out to be reality, our argument in this article would have expelled too many people from the realm of rationality But things are not as they seem Rational ignorance or error and political insincerity account for the behavior of most political self-defeating agents The remaining counterproductive agents are so few that they should not subvert standard accounts of rationality

GUIDO PINCIONE

Universidad Torcuato Di Tella/Buenos Aires

FERNANDO R TESÓN

Arizona State University

---

[23] Nor is there any such constraint in Nozick's account

[24] Is not the consistent agent (see section IV) also insincere? Recall that this agent believes that his support for minimum-wage laws will help the poor, all things considered, yet he also believes that the laws themselves will harm the poor It depends on how one defines insincerity Unlike the posturer, the consistent agent is faithful to the goal or value to which he declares adherence

BOOK REVIEWS

The Nature of Perception. JOHN FOSTER. New York: Oxford University Press, 2000. viii + 289 p. Cloth $55.00.

Perceptual experiences represent that things are thus and so. The experience as of a red, round object two feet in front of me represents that there is something red and round two feet in front of me. I may or may not accept that things are as they are being represented to be; all the same, for each perceptual experience, there is a way that things are being represented to be. As John Foster puts it, there is a view about how things are which the experience invites one to adopt; the invitation may or may not be accepted.

Such a view comes in two forms. *Strong representationalism* holds that the nature of perceptual experience is exhausted in one way or another by its character as a representational state. Give enough detail about the content of how things are being represented to be, and enough detail about the attitude to that content, and you fix the experience without remainder. *Weak representationalism* holds that perceptual experience is essentially representational—no representation, no experience—but that its nature is not exhausted by its representational nature. An experience's representational nature is an essential part of it but a part all the same. Both views are opposed to the view that the connection between experience and representation is contingent and a posteriori, the view that experiences are not essentially representational. The most that is true on this account is that the connection between experience and certain representational states is so well entrenched that it gives the appearance of being a necessary one à la David Hume.

Foster, rightly in my view, finds the third view highly implausible, and sees the debate as being between weak and strong representationalism (though he does not describe them using the weak-strong terminology). He starts by defending a certain way of looking at the debate between these two positions and proceeds to argue for weak representationalism from this perspective. He then turns to making the case for a certain version of weak representationalism. Finally, this version of weak representationalism is argued to lead to idealism. Foster is aware that few will be attracted by the final step and I shall not discuss it further. The interest—the considerable interest—of Foster's book does not depend on finding the final step in any way plausible. I shall be concerned briefly with his discussion of the issue

0022-362X/01/9812/653–57

between the weak and strong representationalism and mostly with his version of the former

Foster argues that the debate between strong and weak representationalism should be seen as concerning the nature of something that can be thought of as *mediating* veridical perceptual contacts Suppose that when I have the experience as of a red, round object two feet in front of me, I have this experience precisely because I am in fact seeing a red, round object two feet in front of me It is natural to think that what constitutes, in this particular case, my seeing the object is the conjunction of an experience which is logically independent of the red, round object I am seeing together with the obtaining of certain facts about the experience's relationship to my environment, including especially the object's role in causing the experience The experience mediates the veridical perceptual contact between what I see and my environment, and is a common element between the veridical perception of the red, round object and various illusions and hallucinations as of red, round objects before me This view is a familiar one but Foster provides a very useful articulation of how it should be understood and defended against those who insist that veridical perception involves what is perceived in a way that does not allow us to "subtract" facts about perceiving subjects' relationships to their environments to get a mediating experience that is the common element in cases of having perceptual experiences, be they veridical, illusory or hallucinatory

The issue for Foster now becomes what to say about the mediating experience—the subjects' side of perception, so to speak Foster's reasons for denying strong representationalism about the mediating experience are familiar (This is not to say that he does not give them new life and force, though, speaking for myself and despite the fact that I once found reasons like them decisive, my present self was not converted by Foster back to the opinions of my previous self) For instance, he argues that no matter how much detail you give about the content of, say, a visual experience, you can always imagine having a state with that very content *C*, perhaps as a result of being told by someone while your eyes are shut that *C* obtains, and yet not having a perceptual experience of the kind in question you can accept something on the basis of testimony without thereby having a visual experience of the relevant kind But if strong representationalism fails, weak representationalists like Foster have the problem of the remainder If experience is essentially representational but not representational without remainder, what is the extra, nonrepresentational element? Foster's response to the failure, as he sees it, of strong representationalism is to offer the *sense-quale theory* (SQT)

Perceptual experiences (of the mediating kind) are a conjunction of a representational part—this ensures that the view is a version of representationalism, and a part which is a relation to a sensory quale or universal. This latter is the nonconceptual sensory core, as he puts it, that strong representationalism leaves out.

SQT arises from a consideration of problems for the traditional *sense-datum theory* (SDT), which has the subject related to a mental item that bears the apparent properties—red, round, or whatever, and the *adverbial theory* (AT), which has the subject sensing in a certain manner—red-ly, round-ly, or whatever. Foster points out that any sense-datum account faces inter alia the problem of how to make sense of an experience, or part of one, being a relation to something which cannot exist independently of the experience. What sort of a *relation* is that? But if we respond by embracing AT, which was of course designed in part to handle this problem, Foster argues that we have inter alia the problem of how to integrate suitably the representational and nonconceptual parts of the total mediating perceptual experience. If, as on SDT, experiences are conjunctions of a conceptual part that represents that things are thus and so, together with a nonconceptual part that is a relation to an item that bears the represented properties (the "thus and so"), we have a version of weak representationalism that makes some sort of sense. For both components are directed to the same set of properties, even if only one component strictly represents them as being instantiated in the world around us. But if the nonrepresentational part does not contain anything that can be thought of as instantiating the represented properties—the round-ly and red-ly ways of sensing that figure in AT are not instances of either roundness or redness—we end up with a theory that is in reality no kind, not even a weak kind, of representationalism. For the truly experiential component of a perceptual experience is not directed to properties that are "pushed out" onto the world around us.

Here is a way of putting what I take to be Foster's essential point against AT. Consider theorists who deny both strong and weak representationalism, holding that there is nothing constitutive or necessary about the connection between perceptual experiences and how things are being represented to be to the subject. Despite their rejection of both strong and weak representationalism, these theorists will (must) allow that, by and large, experiences are accompanied by representational states. Now suppose that they use the term 'experience-*r*' for a conjunction of an experience with an accompanying representational state. How do they differ in any substantive sense from AT theorists? They would appear to be simply using 'experience-*r*' where adverbial theorists use 'experience'. Foster's good thought is that weak representationalists cannot embrace AT. For if they do, the only way they can defend their insistence that experiences have of necessity a representational element is in effect by insisting on using the term 'experience' for the conjunction of a mode of sensing with something intrinsically representational. What was supposed to be an interesting thesis has become a verbal stipulation.

Foster concludes that what we need is an account of the nonconceptual sensory core which analyzes it (1) as a relation to something that exists independently of the experience, so avoiding SDT's big problem, and (2) as a relation to the properties in question, the ones that are being represented to obtain, so avoiding AT's big problem. His solution is to treat the core nonconceptual experience as a relation to a property-like universal, to something that is *one and the same* in all particular cases of having a sensory experience of a certain sort, and not as a relation to different particular objects, sense data, as in the SDT. When I sense red and you sense red, we are related to the very same item—call it a quale, not to two different instances of a red sense datum. At the same time, the quale's existence is not internal to any particular experience. Because the red a subject senses is the very same item as is sensed on all occasions of sensing red, it is not something whose existence depends on any given sensing, though its existence does, according to Foster, depend on being sensed by some subject or other. Although the qualia of this theory are universal-like in that we have one and the same item featuring in different experiences, Foster is explicit that they are not property universals in the sense of universals as abstract objects; qualia are concrete universals, as we might put it. Foster's reason for insisting that they are not abstract universals is that the only way abstract universals "can contribute to the composition of concrete reality" (190) is by instantiation.

I think this is an interesting new theory—well, not entirely new but that is par for the course in philosophy. I close by outlining what seems to me the big problem that faces it. I am not suggesting that Foster is unaware of the problem, though obviously he does not take it to be as serious as I do.

Sense-datum theorists might agree with Foster by stipulation (and it is worth noting that Foster himself talks of his theory as one that employs "a stipulative stroke" (188)). They might use 'quale' for any scattered object which is an aggregation of matching sense data, and they might count you and I as being aware of the very same quale when you and I are aware of sense data that are parts of a single quale as just defined Carter and Nixon were born in the very same country despite being born at different places They were born in different parts of the one country, the United States In the same way, you and I would be aware of numerically the same quale in the stipulated sense by being aware of different parts of the same aggregation Now, this theory would have precisely the two virtues Foster is after First, the nonconceptual core would be a relation to something whose existence is not internal to the experience Aggregations differ from sets inter alia in being able to retain their identity on losing a member So the theory would avoid SDT's big problem Second, the relation would be to something possessing the key properties The parts of the quale are simply the sense data of SDT and they bear the relevant properties So the theory would avoid AT's big problem But *of course* the theory just described is nothing more than the old SDT in new words

I should emphasize that Foster is explicit that his quale are *not* scattered objects (see especially 192-93) But we need an existence proof If they are not property universals as traditionally conceived, and if they are not aggregations of sense data-like instances of the property universals, are "they" anything? Without a good answer, Foster's theory is exposed to the charge of postulating rather than finding a "that which" solves his problem

FRANK JACKSON

Australian National University

INDEX TO VOLUME XCVIII

INDEX OF SUBJECTS DISCUSSED IN THE JOURNAL

Adams, Ernst

Arló-Costa, Horacio Bayesian Epistemology and Epistemic Conditionals On the Status of the Export-Import Laws 98 (11) 555-93 '01

Armstrong, David M.

Clapp, Lenny Disjunctive Properties Multiple Realizations 98 (3) 111-36 '01

Belief

Arló-Costa, Horacio Bayesian Epistemology and Epistemic Conditionals On the Status of the Export-Import Laws 98 (11) 555-93 '01

Everett, Theodore J The Rationality of Science and the Rationality of Faith 98 (1) 19-42 '01

Bayesian Theory

Arló-Costa, Horacio Bayesian Epistemology and Epistemic Conditionals On the Status of the Export-Import Laws 98 (11) 555-93 '01

Biology, Philosophy of

Bradley, Peter Of Colors, Kestrels, Caterpillars, and Leaves By Peter Bradley and Michael Tye 98 (9) 469-87 '01

Tye, Michael Of Colors, Kestrels, Caterpillars, and Leaves By Peter Bradley and Michael Tye 98 (9) 469-87 '01

Causation

Hitchcock, Christopher The Intransitivity of Causation Revealed in Equations and Graphs 98 (6) 273-99 '01

Schaffer, Jonathan Causes as Probability Raisers of Processes 98 (2) 75-92 '01

Cognitive Science

Blackmon, James Systematicity and the Cognition of Structured Domains By Robert Cummins, James Blackmon, David Byrd, Pierre Poirier, Martin Roth, and Georg Schwarz 98 (4) 167-85 '01

Byrd, David Systematicity and the Cognition of Structured Domains By Robert Cummins, James Blackmon, David Byrd, Pierre Poirier, Martin Roth, and Georg Schwarz 98 (4) 167-85 '01

Cummins, Robert Systematicity and the Cognition of Structured Domains By Robert Cummins, James Blackmon, David Byrd, Pierre Poirier, Martin Roth, and Georg Schwarz 98 (4) 167-85 '01

Giaquinto, Marcus Knowing Numbers 98 (1) 5-18 '01

Rupert, Robert D Coining Terms in the Language of Thought Innateness, Emergence, and the Lot of Cummins's Argument against the Causal Theory 98 (10) 499-530 '01

Poirier, Pierre Systematicity and the Cognition of Structured Domains By Robert Cummins, James Blackmon, David Byrd, Pierre Poirier, Martin Roth, and Georg Schwarz 98 (4) 167-85 '01

Roth, Martin Systematicity and the Cognition of Structured Domains By Robert Cummins, James Blackmon, David Byrd, Pierre Poirier, Martin Roth, and Georg Schwarz 98 (4) 167-85 '01

Schwarz, Georg Systematicity and the Cognition of Structured Domains By Robert Cummins, James Blackmon, David Byrd, Pierre Poirier, Martin Roth, and Georg Schwarz 98 (4) 167-85 '01

Colors

Bradley, Peter Of Colors, Kestrels, Caterpillars, and Leaves By Peter Bradley and Michael Tye **98** (9) 469-87 '01

Tye, Michael Of Colors, Kestrels, Caterpillars, and Leaves By Peter Bradley and Michael Tye **98** (9) 469-87 '01

Counterfactuals

Schaffer, Jonathan Causes as Probability Raisers of Processes **98** (2) 75-92 '01

Cummins, Robert

Rupert, Robert D Coining Terms in the Language of Thought Innateness, Emergence, and the Lot of Cummins's Argument against the Causal Theory **98** (10) 499-530 '01

Davidson, Donald

Kölbel, Max Two Dogmas of Davidsonian Semantics **98** (12) 613-35 '01

Einstein, Albert

Lange, Marc The Most Famous Equation **98** (5) 219-38 '01

Epistemology: *See* Knowledge, Theory of

Ethics

Hubin, Donald C The Groundless Normativity of Instrumental Rationality **98** (9) 445-68 '01

Explanation

Hitchcock, Christopher The Intransitivity of Causation Revealed in Equations and Graphs **98** (6) 273-99 '01

Fodor, Jerry

Clapp, Lenny Disjunctive Properties Multiple Realizations **98** (3) 111-36 '01

Frege, Gottlob

Justice, John On Sense and Reflexivity **98** (7) 351-64 '01

Game Theory

Brams, Steven J Paradoxes of Fair Division By Steven J Brams, Paul H Edelman, and Peter C Fishburn **98** (6) 300-14 '01

Edelman, Paul H Paradoxes of Fair Division By Steven J Brams, Paul H Edelman, and Peter C Fishburn **98** (6) 300-14 '01

Fishburn, Peter C Paradoxes of Fair Division By Steven J Brams, Paul H Edelman, and Peter C Fishburn **98** (6) 300-14 '01

Good Life

Wall, Steven Neutrality and Responsibility **98** (8) 389-410 '01

Hacking, Ian

Belot, Gordon The Principle of Sufficient Reason **98** (2) 55-74 '01

Hume, David

Hubin, Donald C The Groundless Normativity of Instrumental Rationality **98** (9) 445-68 '01

Justice

Brams, Steven J Paradoxes of Fair Division By Steven J Brams, Paul H Edelman, and Peter C Fishburn **98** (6) 300-14 '01

Edelman, Paul H Paradoxes of Fair Division By Steven J Brams, Paul H Edelman, and Peter C Fishburn **98** (6) 300-14 '01

Fishburn, Peter C Paradoxes of Fair Division By Steven J Brams, Paul H Edelman, and Peter C Fishburn **98** (6) 300-14 '01

Wall, Steven Neutrality and Responsibility **98** (8) 389-410 '01

Kim, Jaegwon

Clapp, Lenny Disjunctive Properties Multiple Realizations **98** (3) 111-36 '01

Knowledge, Theory of

Arló-Costa, Horacio Bayesian Epistemology and Epistemic Conditionals On the Status of the Export-Import Laws **98** (11) 555-93 '01

Everett, Theodore J The Rationality of Science and the Rationality of Faith **98** (1) 19-42 '01

Giaquinto, Marcus Knowing Numbers **98** (1) 5-18 '01

Peacocke, Christopher Does Perception Have a Nonconceptual Content? **98** (5) 239-64 '01

Putnam, Hilary When "Evidence Transcendence" Is Not Malign A Reply to Crispin Wright **98** (11) 594-600 '01

Schaffer, Jonathan Causes as Probability Raisers of Processes **98** (2) 75-92 '01

Stanley, Jason Knowing How By Jason Stanley and Timothy Williamson **98** (8) 411-44 '01

Williamson, Timothy Knowing How By Jason Stanley and Timothy Williamson **98** (8) 411-44 '01

Kripke, Saul

Justice, John On Sense and Reflexivity **98** (7) 351-64 '01

Katz, Jerrold J The End of Millianism Multiple Bearers, Improper Names, and Compositional Meaning **98** (3) 137-66 '01

Language, Philosophy of

Justice, John On Sense and Reflexivity **98** (7) 351-64 '01

Katz, Jerrold J The End of Millianism Multiple Bearers, Improper Names, and Compositional Meaning **98** (3) 137-66 '01

Kolbel, Max Two Dogmas of Davidsonian Semantics **98** (12) 613-35 '01

Pitt, David Alter Egos and Their Names **98** (10) 531-52 '01

Leibniz, Gottfried

Belot, Gordon The Principle of Sufficient Reason **98** (2) 55-74 '01

Lewis, David

Hitchcock, Christopher The Intransitivity of Causation Revealed in Equations and Graphs **98** (6) 273-99 '01

Schaffer, Jonathan Causes as Probability Raisers of Processes **98** (2) 75-92 '01

Mathematics, Philosophy of

Bays, Timothy On Putnam and His Methods **98** (6) 331-50 '01

Giaquinto, Marcus Knowing Numbers **98** (1) 5-18 '01

Meaning

Kolbel, Max Two Dogmas of Davidsonian Semantics **98** (12) 613-35 '01

Memo to Authors

98 (2) 101 '01

Memory

Slors, Marc Personal Identity, Memory, and Circularity An Alternative for Q-Memory **98** (4) 186-214 '01

Metaphysics

Belot, Gordon The Principle of Sufficient Reason **98** (2) 55-74 '01

Clapp, Lenny Disjunctive Properties Multiple Realizations **98** (3) 111-36 '01

Hitchcock, Christopher The Intransitivity of Causation Revealed in Equations and Graphs **98** (6) 273-99 '01

Pitt, David Alter Egos and Their Names **98** (10) 531-52 '01

Sllors, Marc Personal Identity, Memory, and Circularity An Alternative for Q-Memory **98** (4) 186-214 '01

Mill, John Stuart

Katz, Jerrold J The End of Millianism Multiple Bearers, Improper Names, and Compositional Meaning **98** (3) 137-66 '01

Mind, Philosophy of

Blackmon, James Systematicity and the Cognition of Structured Domains By Robert Cummins, James Blackmon, David Byrd, Pierre Poirier, Martin Roth, and Georg Schwarz **98** (4) 167-85 '01

Byrd, David Systematicity and the Cognition of Structured Domains By Robert Cummins, James Blackmon, David Byrd, Pierre Poirier, Martin Roth, and Georg Schwarz **98** (4) 167-85 '01

Clapp, Lenny Disjunctive Properties Multiple Realizations **98** (3) 111-36 '01

Cummins, Robert Systematicity and the Cognition of Structured Domains By Robert Cummins, James Blackmon, David Byrd, Pierre Poirier, Martin Roth, and Georg Schwarz **98** (4) 167-85 '01

Peacocke, Christopher Does Perception Have a Nonconceptual Content? **98** (5) 239-64 '01

Poirier, Pierre Systematicity and the Cognition of Structured Domains By Robert Cummins, James Blackmon, David Byrd, Pierre Poirier, Martin Roth, and Georg Schwarz **98** (4) 167-85 '01

Rosenberg, Alex On Multiple Realization and the Special Sciences **98** (7) 365-73 '01

Roth, Martin Systematicity and the Cognition of Structured Domains By Robert Cummins, James Blackmon, David Byrd, Pierre Poirier, Martin Roth, and Georg Schwarz **98** (4) 167-85 '01

Rupert, Robert D Coining Terms in the Language of Thought Innateness, Emergence, and the Lot of Cummins's Argument against the Causal Theory **98** (10) 499-530 '01

Schwarz, Georg Systematicity and the Cognition of Structured Domains By Robert Cummins, James Blackmon, David Byrd, Pierre Poirier, Martin Roth, and Georg Schwarz **98** (4) 167-85 '01

Slors, Marc Personal Identity, Memory, and Circularity An Alternative for Q-Memory **98** (4) 186-214 '01

Multiple Realizability Thesis

Clapp, Lenny Disjunctive Properties Multiple Realizations **98** (3) 111-36 '01

New Books

98 (2) 102-10 '01

98 (7) 379-88 '01

New Books Anthologies

98 (4) 215-17 '01

98 (10) 553-54 '01

New Books Translations

98(4) 218 '01

Nominalism

Giaquinto, Marcus Knowing Numbers **98** (1) 5-18 '01

Notes and News

98 (6) 330 '01

98 (9) 498 '01

Normativity

Hubin, Donald C The Groundless Normativity of Instrumental Rationality **98** (9) 445-68 '01

Nozick, Robert

Pincione, Guido Self-Defeating Symbolism in Politics By Guido Pincione and Fernando R Tesón **98** (12) 636-52 '01

Tesón, Fernando R Self-Defeating Symbolism in Politics By Guido Pincione and Fernando R Tesón **98** (12) 636-52 '01

Numbers

Giaquinto, Marcus Knowing Numbers **98** (1) 5-18 '01

Paradoxes

Brams, Steven J Paradoxes of Fair Division By Steven J Brams, Paul H Edelman, and Peter C Fishburn **98** (6) 300-14 '01

Edelman, Paul H Paradoxes of Fair Division By Steven J Brams, Paul H Edelman, and Peter C Fishburn **98** (6) 300-14 '01

Fishburn, Peter C Paradoxes of Fair Division By Steven J Brams, Paul H Edelman, and Peter C Fishburn **98** (6) 300-14 '01

Perception

Bradley, Peter Of Colors, Kestrels, Caterpillars, and Leaves By Peter Bradley and Michael Tye **98** (9) 469-87 '01

Peacocke, Christopher Does Perception Have a Nonconceptual Content? **98** (5) 239-64 '01

Tye, Michael Of Colors, Kestrels, Caterpillars, and Leaves By Peter Bradley and Michael Tye **98** (9) 469-87 '01

Personal Identity

Slors, Marc Personal Identity, Memory, and Circularity An Alternative for Q-Memory **98** (4) 186-214 '01

Philosophy, Political· *See* Political Theory

Physicalism

Clapp, Lenny Disjunctive Properties Multiple Realizations **98** (3) 111-36 '01

Physics, Philosophy of

Lange, Marc The Most Famous Equation **98** (5) 219-38 '01

Political Theory

Brams, Steven J Paradoxes of Fair Division By Steven J Brams, Paul H Edelman, and Peter C Fishburn **98** (6) 300-14 '01

Edelman, Paul H Paradoxes of Fair Division By Steven J Brams, Paul H Edelman, and Peter C Fishburn **98** (6) 300-14 '01

Fishburn, Peter C Paradoxes of Fair Division By Steven J Brams, Paul H Edelman, and Peter C Fishburn **98** (6) 300-14 '01

Pincione, Guido Self-Defeating Symbolism in Politics By Guido Pincione and Fernando R Tesón **98** (12) 636-52 '01

Tesón, Fernando R Self-Defeating Symbolism in Politics By Guido Pincione and Fernando R Tesón **98** (12) 636-52 '01

Wall, Steven Neutrality and Responsibility **98** (8) 389-410 '01

Probability

Arló-Costa, Horacio Bayesian Epistemology and Epistemic Conditionals On the Status of the Export-Import Laws **98** (11) 555-93 '01

Schaffer, Jonathan Causes as Probability Raisers of Processes **98** (2) 75-92 '01

Psychology, Philosophy of

Peacocke, Christopher Does Perception Have a Nonconceptual Content? **98** (5) 239-64 '01

Rosenberg, Alex On Multiple Realization and the Special Sciences **98** (7) 365-73 '01

Slors, Marc Personal Identity, Memory, and Circularity An Alternative for Q-Memory **98** (4) 186-214 '01

Putnam, Hilary

Bays, Timothy On Putnam and His Methods **98** (6) 331-50 '01

Rosenberg, Alex On Multiple Realization and the Special Sciences **98** (7) 365-73 '01

Rationality

Everett, Theodore J The Rationality of Science and the Rationality of Faith **98** (1) 19-42 '01

Hubin, Donald C The Groundless Normativity of Instrumental Rationality **98** (9) 445-68 '01

Pincione, Guido Self-Defeating Symbolism in Politics By Guido Pincione and Fernando R Tesón **98** (12) 636-52 '01

Tesón, Fernando R Self-Defeating Symbolism in Politics By Guido Pincione and Fernando R Tesón **98** (12) 636-52 '01

Realism

Bays, Timothy On Putnam and His Methods 98 (6) 331-50 '01

Bradley, Peter Of Colors, Kestrels, Caterpillars, and Leaves By Peter Bradley and Michael Tye 98 (9) 469-87 '01

Putnam, Hilary When "Evidence Transcendence" Is Not Malign A Reply to Crispin Wright 98 (11) 594-600 '01

Tye, Michael Of Colors, Kestrels, Caterpillars, and Leaves By Peter Bradley and Michael Tye 98 (9) 469-87 '01

Reference

Pitt, David Alter Egos and Their Names 98 (10) 531-52 '01

Religion, Philosophy of

Everett, Theodore J The Rationality of Science and the Rationality of Faith 98 (1) 19-42 '01

Representation

Rupert, Robert D Coining Terms in the Language of Thought Innateness, Emergence, and the Lot of Cummins's Argument against the Causal Theory 98 (10) 499-530 '01

Ryle

Stanley, Jason Knowing How By Jason Stanley and Timothy Williamson 98 (8) 411-44 '01

Williamson, Timothy Knowing How By Jason Stanley and Timothy Williamson 98 (8) 411-44 '01

Science, Philosophy of

Belot, Gordon The Principle of Sufficient Reason 98 (2) 55-74 '01

Everett, Theodore J The Rationality of Science and the Rationality of Faith 98 (1) 19-42 '01

Giaquinto, Marcus Knowing Numbers 98 (1) 5-18 '01

Lange, Marc The Most Famous Equation 98 (5) 219-38 '01

Rosenberg, Alex On Multiple Realization and the Special Sciences 98 (7) 365-73 '01

Semantics

Kolbel, Max Two Dogmas of Davidsonian Semantics 98 (12) 613-35 '01

Pitt, David Alter Egos and Their Names 98 (10) 531-52 '01

Shapiro, Lawrence

Rosenberg, Alex On Multiple Realization and the Special Sciences 98 (7) 365-73 '01

Symbolism

Pincione, Guido Self-Defeating Symbolism in Politics By Guido Pincione and Fernando R Tesón 98 (12) 636-52 '01

Tesón, Fernando R Self-Defeating Symbolism in Politics By Guido Pincione and Fernando R Tesón 98 (12) 636-52 '01

Tarski, Alfred

Kölbel, Max Two Dogmas of Davidsonian Semantics 98 (12) 613-35 '01

Testimony

Everett, Theodore J The Rationality of Science and the Rationality of Faith 98 (1) 19-42 '01

Truth

Kölbel, Max Two Dogmas of Davidsonian Semantics 98 (12) 613-35 '01

Putnam, Hilary When "Evidence Transcendence" Is Not Malign A Reply to Crispin Wright 98 (11) 594-600 '01

Wright, Crispin

Putnam, Hilary When "Evidence Transcendence" Is Not Malign A Reply to Crispin Wright 98 (11) 594-600 '01

INDEX OF AUTHORS

Including Contributing Authors and Reviewers

Arló-Costa, Horacio, Bayesian Epistemology and Epistemic Conditionals On the Status of the Export-import Laws **98** (11) 555-93 '01

Bays, Timothy, On Putnam and His Models **98** (6) 331-50 '01
Belot, Gordon, The Principle of Sufficient Reason **98** (2) 55-74 '01
Blackmon, James, Systematicity and the Cognition of Structured Domains By Robert Cummins, James Blackmon, David Byrd, Pierre Poirier, Martin Roth, and Georg Schwarz **98** (4) 167-85 '01
Bradley, Peter, Of Colors, Kestrels, Caterpillars, and Leaves By Peter Bradley and Michael Tye **98** (9) 469-87 '01
Brams, Steven J , Paradoxes of Fair Division By Steven J Brams, Paul H Edelman, and Peter C Fishburn **98** (6) 300-14 '01
Byrd, David, Systematicity and the Cognition of Structured Domains By Robert Cummins, James Blackmon, David Byrd, Pierre Poirier, Martin Roth, and Georg Schwarz **98** (4) 167-85 '01

Clapp, Lenny, Disjunctive Properties Multiple Realizations **98** (3) 111-36 '01
Cummins, Robert, Systematicity and the Cognition of Structured Domains By Robert Cummins, James Blackmon, David Byrd, Pierre Poirier, Martin Roth, and Georg Schwarz **98** (4) 167-85 '01

Edelman, Paul H , Paradoxes of Fair Division By Steven J Brams, Paul H Edelman, and Peter C Fishburn **98** (6) 300-14 '01
Everett, Theodore J , The Rationality of Science and the Rationality of Faith **98** (1) 19-42 '01

Fishburn, Peter C , Paradoxes of Fair Division By Steven J Brams, Paul H Edelman, and Peter C Fishburn **98** (6) 300-14 '01

Giaquinto, Marcus, Knowing Numbers **98** (1) 5-18 '01
Goodin, Robert E , Review of Russell Hardin *Liberalism, Constitutionalism, and Democracy* **98** (7) 374-78 '01

Harman, Gilbert, Review of Noam Chomsky *New Horizons in the Study of Language and Mind* **98** (5) 265-69 '01
Hill, Thomas E , Review of John Rawls *Collected Papers* **98** (5) 269-72 '01
Hitchcock, Christopher, The Intransitivity of Causation Revealed in Equations and Graphs **98** (6) 273-99 '01
Hubin, Donald C , The Groundless Normativity of Instrumental Rationality **98** (9) 445-68 '01

Jackson, Frank, Review of John Foster *The Nature of Perception* **98** (12) 653-57 '01
Justice, John, On Sense and Reflexivity **98** (7) 351-64 '01

Katz, Jerrold J , The End of Millianism Multiple Bearers, Improper Names, and Compositional Meaning **98** (3) 137-66 '01
Kolbel, Max, Two Dogmas of Davidsonian Semantics **98** (12) 613-35 '01

Loewer, Barry, Review of Jaegwon Kim *Mind in a Physical World Essay on the Mind-body Problem and Mental Causation* **98** (6) 315-34 '01
Lange, Marc, The Most Famous Equation **98** (5) 219-38 '01

Markosian, Ned, Review of Peter Ludlow *Semantics, Tense, and Time* **98** (6) 325-29 '01
McKenna, Michael, Review of John Martin Fischer and Mark Ravizza *Responsibility and Control A Theory of Moral Responsibility* **98** (2) 93-100 '01

Peacocke, Christopher, Does Perception Have a Nonconceptual Content? **98** (5) 239-64 '01
Pincione, Guido, Self-Defeating Symbolism in Politics By Guido Pincione

and Fernando R Tesón 98 (12) 635-52 '01

Pitt, David, Alter Egos and Their Names 98 (10) 531-52 '01

Poirier, Pierre, Systematicity and the Cognition of Structured Domains By Robert Cummins, James Blackmon, David Byrd, Pierre Poirier, Martin Roth, and Georg Schwarz 98 (4) 167-85 '01

Putnam, Hilary, When "Evidence Transcendence" Is Not Malign A Reply to Crispin Wright 98 (11) 594-600 '01

Rosenberg, Alex, On Multiple Realization and the Special Sciences 98 (7) 365-73 '01

Roth, Martin, Systematicity and the Cognition of Structured Domains By Robert Cummins, James Blackmon, David Byrd, Pierre Poirier, Martin Roth, and Georg Schwarz 98 (4) 167-85 '01

Rupert, Robert D , Coining Terms in the Language of Thought Innateness, Emergence, and the Lot of Cummins's Argument against the Causal Theory 98 (10) 499-530 '01

Schaffer, Jonathan, Causes as Probability Raisers of Processes 98 (2) 75-92 '01

Schwarz, Georg, Systematicity and the Cognition of Structured Domains By Robert Cummins, James Blackmon, David Byrd, Pierre Poirier, Martin Roth, and Georg Schwarz 98 (4) 167-85 '01

Slors, Marc, Personal Identity, Memory, and Circularity An Alternative for Q-Memory 98 (4) 186-214 '01

Sluga, Hans, Review of Michael Friedman *A Parting of the Ways Carnap, Cassirer, and Heidegger* 98 (11) 601-11 '01

Stanford, P. Kyle, Review of Elliott Sober and David Sloan Wilson *Unto Others The Evolution and Psychology of Unselfish Behavior* 98 (1) 43-47 '01

Stanley, Jason, Knowing How By Jason Stanley and Timothy Williamson 98 (8) 411-44 '01

Tappenden, Jamie, Recent Work in Philosophy of Mathematics 98 (9) 488-97 '01

Tesón, Fernando R., Self-Defeating Symbolism in Politics By Guido Pincione and Fernando R Tesón 98 (12) 635-52 '01

Tye, Michael, Of Colors, Kestrels, Caterpillars, and Leaves By Peter Bradley and Michael Tye 98 (9) 469-87 '01

Wall, Steven, Neutrality and Responsibility 98 (8) 389-410 '01

Williamson, Timothy, Knowing How By Jason Stanley and Timothy Williamson 98 (8) 411-44 '01

Wilson, Catherine, Review of Carol Rovane *The Bounds of Agency An Essay in Revisionary Metaphysics* 98 (1) 47-54 '01

INDEX OF BOOKS REVIEWED

Listed by Author of Item Reviewed

Chomsky, Noam. *New Horizons in the Study of Language and Mind*—Gilbert Harman 98 (5) 265-69 '01

Fischer, John Martin, and **Mark Ravizza** *Responsibility and Control A Theory of Moral Responsibility*—Michael McKenna 98 (2) 93-100 '01

Friedman, Michael *A Parting of the Ways Carnap, Cassirer, and Heidegger*—Hans Sluga 98 (11) 601-11 '01

Foster, John *The Nature of Perception*—Frank Jackson 98 (12) 653-57 '01

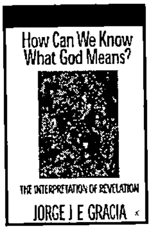
Hardin, Russell· *Liberalism, Constitutionalism, and Democracy*—Robert E Goodin 98 (7) 374-78 '01

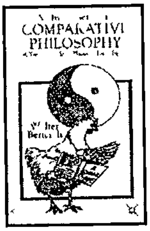
Kim, Jaegwon *Mind in a Physical World Essay on the Mind-body Problem and Mental Causation*—Barry Loewer 98 (6) 315-34 '01

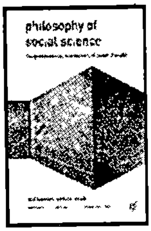
Ludlow, Peter· *Semantics, Tense, and Time*—Ned Markosian 98 (6) 325-29 '01

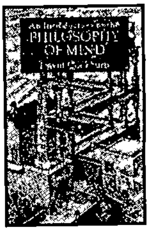
Maddy, Penelope *Naturalism in Mathemathics*—Jamie Tappenden 98 (9) 488-97 '01

Ravizza, Mark, and **John Martin Fischer:** *Responsibility and Control A Theory of Moral Responsibility*—Michael McKenna **98** (2) 93-100 '01
Rawls, John *Collected Papers*—Thomas E Hill **98** (5) 269-72 '01
Resnik, Michael *Mathematics as a Science of Patterns*—Jamie Tappenden **98** (9) 488-97 '01
Rovane, Carol *The Bounds of Agency An Essay in Revisionary Metaphysics*—Catherine Wilson **98** (1) 47-54 '01
Shapiro, Stewart *Philosophy of Mathematics Structure and Ontology*—Jamie Tappenden **98** (9) 488-97 '01
Sober, Elliot, and **David Sloan Wilson** *Unto Others The Evolution and Psychology of Unselfish Behavior*—P Kyle Stanford **98** (1) 43-47 '01
Wilson, David Sloan, and **Elliott Sober** *Unto Others The Evolution and Psychology of Unselfish Behavior*—P Kyle Stanford **98** (1) 43-47 '01


How Can We Know What God Means?
THE INTERPRETATION OF REVELATION
JORGE J E GRACIA







How Can We Know What God Means?
The Interpretation of Revelation
Jorge J E Gracia
256 pp / 0-312-24028-7 / $18 95 pb

Introduction to Comparative Philosophy
A Travel Guide to Philosophical Space
Walter Benesch
239 pp / 0-333-93068-1 / $19 95 pb

In Defense of Sin
Edited by John Portman
320 pp / 0-312-23986-6 / $23 95 cl

Philosophy of Social Science
Philosophical Issues in Social Thought
Ted Benton and Ian Craib
224 pp / 0-333-77499-X / $19 95 pb

An Introduction to the Philosophy of Mind
David Cockburn
177 pp / 0-333-96122-6 / $18 95 pb

Marx@2000
Late Marxist Perspectives
Ronaldo Munck
176 pp / 1-84277-083-7 / $19 95 pb

palgrave *Global Publishing at St Martin's Press*
1-800-221-7945 ext 270 • www palgrave-usa com
Distributors of I B Tauris, Manchester University Press and Zed Books

REVUE INTERNATIONALE DE PHILOSOPHIE

Editor: Michel MEYER
Université de Bruxelles, C.P 188
50, avenue F.D. Roosevelt - 1050 Brussels, Belgium

Each number is devoted to a particular movement, a particular philosopher, or a particular problem.

We publish 4 issues annually Articles are written in English, French, German, or Italian

Our last issue was devoted to *Searle with his replies*

Articles by: Daniel Vanderveken, *Présentation* – John R. Searle, *Meaning, Mind and Reality* – Hubert L Dreyfus, *Phenomenological Description versus Rational Reconstruction* – François Recanati, *Déstabiliser le sens* – Dietmar Zaefferer, *Deconstructing a Classical Classification A Typological Look at Searle's Concept of Illocution Type* – Francis Jacques, *Remarques sur la promesse et le pardon La Théorie des actes de langage à l'épreuve de l'éthique* – Daniel Vanderveken, *Illocutionary Logic and Discourse Typology* — Analyses critiques: David Sosa, *Slouching Towards Dualism* – Barry Barnes, *On the Construction of Social Reality* – André Leclerc, *From the Mystery to the Problem of Consciousness* — Réponses de Searle

Coming numbers Bourdieu et la philosophie – L'Esthétique de Hegel – Sociologie des Sciences – Ryle – Le Corps – Putnam – Philosophy of Economics – Wittgenstein on Esthetics and Ethics – Dilthey

Cambridge Journals

Forthcoming
Decisions and Games - Special Issue in 2002

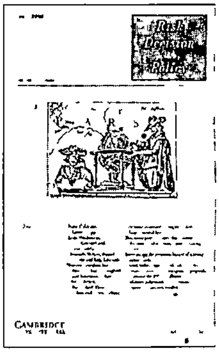
Risk, Decision and Policy

The official journal of the Decision and Policy Network

Established in 1994, *Risk Decision and Policy* is an interdisciplinary journal with an emphasis on decision and risk research. The journal considers applications of descriptive, prescriptive and normative decision research in a wide variety of areas. These could range from political concerns (such as voting issues, and collective choice problems) to peace and security, to long-term sustainability, to issues of interest to the private sector, such as R & D. The central core is an emphasis on decision research as a means to useful insight for practical issues, drawing on theory and applications. The journal will be particularly valuable to social scientists and economists whose work might also be of relevance to policy-makers in international and national organisations, especially in the health and environment areas. As well as technical and occasional survey articles, the journal carries regular symposia edited by Guest Editors which cover a variety of topics including AIDS, BSE, Imprecise Probabilities and Risk Communication.

Subscriptions

Volume 6 in 2001: April, June and December
Institutions, print and electronic: £82/$135
Individuals print only: £30/$47
The Decision Analysis Society, European Association of Decisions Making, Society for Risk Analysis, Risk Assessment & Policy Association, Association for Environmental and Resource Economists, European Association for Environmental and Resource Economists: £24/$39
ISSN 1357-5309



Please feel free to visit the journal homepage at
http://journals.cambridge.org/journal_riskdecisionandpolicy

to contact the Customer Services Journals Division –
in Cambridge *tel* +44 (0)1223 326070 *fax* +44 (0)1223 325150 *email* journals_subscriptions@cambridge.org
in New York *tel* (914) 937 9600 *fax* (914) 937 4712 *email* journals_subscriptions@cup.org

CAMBRIDGE
UNIVERSITY PRESS

The Edinburgh Building, Cambridge, CB2 2RU, UK
40 West 20th Street, New York, NY 10011-4211, USA